Public Health & Intelligence

PHI Trend Analysis Guidance

Document Control	
Version	Version 1.5
Date Issued	March 2017
Author	Róisín Farrell, David Walker / HI Team
Comments to	NSS.isdsubstancemisuse@nhs.net

Version	Date	Comment	Author
Version 1.0	July 2016	1st version of paper	David Walker, Róisín Farrell
Version 1.1	August 2016	Amending format; adding sections	Róisín Farrell
Version 1.2	November 2016	Adding content to sections 1 and 2	Róisín Farrell, Lee Barnsdale
Version 1.3	November 2016	Amending as per suggestions from SAG	Róisín Farrell
Version 1.4	December 2016	Structural changes	Róisín Farrell, Lee Barnsdale
Version 1.5	March 2017	Amending as per suggestions from SAG	Róisín Farrell, Lee Barnsdale

Contents

1. Introduction

When presenting and publishing data, it is common to provide a trend or time series analysis. A time series is simply a sequence of data points plotted over time (frequently using column or line charts). Time series analysis is used for a number of reasons:

- To summarise a trend and show if a measure is increasing or decreasing. By summarising data across a range of years, it may be possible to remove the 'noise' of a single-year analysis and expose an underlying trend.
- To project future trends, or estimate uncertain past events.
- To identify a change in trend resulting from policy change or a significant event.

In many cases, a description of the observations and the magnitude of any observed change over time may be sufficient. However, when change occurs (and in some cases, when it does not) we may be required to explain why this has happened.

This document provides individuals producing analytical outputs with guidance on time trend analysis methods and examples of their appropriate use. It first examines methods of describing trend data in order to aid interpretation and then discusses methods for exploring associations within data over time, observing the key principle that **the best method to use is the simplest one for the task**.

Examples of current practices within PHI publications are included in Appendix 1.

Worked SPSS/Excel examples of methods described in this paper are presented in Appendix 2. Please note that other statistical packages may produce different results from those stated in this guidance.

2. Understanding trend data for descriptive analysis

Before performing any analysis, it is essential to understand your data. Familiarise yourself with the data collection, submission and validation processes:

- How complete are the data?
- Who assesses completion and how?
- Are there any known data collection/submission issues?
- How are the data validated and quality checked?
- Who carries out validation/data quality checks?
- What is the scope and frequency of validation/data quality checks?
- Are there any known data quality issues?

Issues identified at this stage (e.g. changes in personnel, coding changes, non-submission) may help explain observed changes in indicators. However, in the absence of any issues associated with the collection and management of the data, observed patterns may merit further investigation.

Basic methods of examining your data include performing **frequencies** to describe the observations and **crosstabulations** to identify potential associations between variables (essentially, trend analyses are crosstabulations by time). Using these analyses, you may notice trends such as seasonality effects, where data shows a recurring change at a specific point each year or periods in which changes appear to have occurred. In those instances, your knowledge of the dataset may adequately explain such changes, or it may be appropriate to investigate further.

3. Descriptive analysis of time trends

Having investigated the structure of your data, it will often be necessary to describe the observed trend. The techniques discussed below are methods of presenting the data to identify or highlight the key messages. They can help to answer important questions about outlying data points and inter-year variation (smoothing), changes over time (population proportion testing) or help provide an alternative perspective on trend data (logarithmic scaling).

Most routine PHI publications include trend analyses. However, many are descriptive and do not use sophisticated methods to identify trends in data. Examples of descriptive analysis discussed further in Appendix 1 are:

- ScotPHO Profiles – many of the Profiles available on ScotPHO include simple smoothing in the form of 2-year, 3-year or 5-year rolling averages.
- Drug Related Deaths (DRD) Report – this publication uses population proportions to determine if differences between years are significant.

3.1 Smoothing

Smoothing is a method used in descriptive statistics to help overcome inter-year variation across a time series. This analysis can be performed in Excel or SPSS using simple arithmetic functions (see Appendix 2 for examples of SPSS syntax).

By reducing the impact of inter-year variation, the methods described below may help to produce more robust figures than a simple trend. However, the resulting analysis may not be strictly accurate or reliable. The main reason for this is that data are unweighted (each data point included in the calculation of the rolling average has equal importance in the equation). Therefore, before deciding which of the following methods to use, consider the standard deviation[1] of the data. If the standard deviation is large, median smoothing may be the most appropriate method; otherwise simple smoothing should be sufficient.

3.1.1 Simple smoothing

Simple smoothing is commonly used with time series data to smooth out short term irregularities (peaks and troughs), allowing identification of longer term trends. The simplest example of this is a **simple moving average** (or rolling average), where the mean average of the data from successive years, often three or five, is plotted instead of, or in addition to, the data points.

For an example of simple smoothing in a PHI publication, please see Appendix 1. Example SPSS syntax and a comparison of single year and smoothed trends can be found in Appendix 2.

3.1.2 Median smoothing

Median smoothing is considered more robust than simple smoothing where there is considerable variation from the mean. This process **identifies the median in a list of values** (instead of the mean) over a specified time period. As with simple smoothing, data of a specified time series (e.g. three or five successive years) is used to calculate a median value. It is a more appropriate method to use median smoothing where the underlying data contains outliers which may influence the mean.

[1] Standard deviation is a measure of variation in a set of data values. A low standard deviation implies that there is little deviation from the mean of the set of values, while a high standard deviation indicates that the values encompass a wider range (Wikipedia: accessed December 19th 2016).

3.2 Measuring change

3.2.1 Confidence Intervals

A **Confidence Interval** is a range of values that is normally used to describe the uncertainty around a point estimate of a quantity (for example, a mortality rate). For indicators based on a sample of a population, uncertainty arises from differences between the sample and the population itself. The values of the confidence intervals are therefore considered to be an estimate of the range of true or 'underlying' values. Confidence intervals quantify the uncertainty in this estimate and, generally speaking, describe how different the point estimate could have been if the underlying conditions stayed the same, but chance had led to a different set of data. The wider the confidence interval is, the greater the uncertainty in the estimate. Confidence intervals of 95% are used most often. This means that there is a 19 in 20 chance that the confidence interval holds the 'true' value. The use of 95% is somewhat arbitrary, but is conventional practice in medical and public health statistics. As with standard deviations, large confidence intervals may mean that your data aren't representative of the population.

An example of confidence intervals in a publication can be seen in the ScotPHO profiles in Appendix 1. Confidence Intervals are normally represented on charts by the use of vertical lines extending above and below each data point to represent their upper and lower limits. There are different ways of calculating confidence intervals, depending on whether the indicator used is a proportion or an average[2].

In terms of trend analysis, if the range of values between the upper and lower confidence interval limits of a data point does not overlap with the confidence interval range of a comparison data point, it can be regarded as different.

3.2.2 Population Proportion Tests

When a change is observed in trend data, it is often important to determine if the change was significant. One method for doing this is by evaluating the change in population proportion across years (e.g. from first year in series to most recent year, or from previous year to most recent year).

Population proportion tests estimate the size of the difference in an indicator between two populations. For example, if you wanted to determine if an observed reduction in the number of stillbirths was significant, this method could provide an answer using data on the number of events (stillbirths) and the size of the population (total number of term pregnancies). This test can be conducted within Excel or in SPSS. A full explanation of this method can be found in Appendix 2.

3.2.3 Comparing standardised rates

It may also be beneficial to explore changes in standardised rates, in order to determine if observed differences are significant. While there is no specific test for doing so, a standard incidence ratio can help to determine the significance of the difference between two points in time. Public Health England has released a useful Excel spreadsheet [3] for this purpose, where you input the values from your data and significance calculations are automatically performed.

---

[2] For an example of how to calculate confidence intervals using each method, see Khan Academy's tutorial (accessed December 19th 2016)

[3] Found at http://www.apho.org.uk/resource/view.aspx?RID=48617 [opens Excel spreadsheet]

However, this method is only useful if you're looking just at a start point and an end point, instead of many points within a trend. If you wish to determine if the overall trend is changing significantly, a Poisson regression (described in further detail below) may be the most appropriate method.

3.2.4 Statistical Process Control

Statistical Process Control (SPC)[4] techniques can also be used to highlight areas for further investigation. SPC outputs were developed in industry using measurements from routine processes to establish the number and extent of deviations from an average or to quantify nonconformities over time. SPC techniques have been utilised within healthcare settings to monitor and address numbers of adverse events etc. It is not within scope for this guidance to provide a comprehensive account of SPC, but the main characteristics of 'Control' and 'Run' charts are summarised below.

'Control' charts can be used to detect deviations from the mean over time using control limits (usually ±3 standard deviations (equivalent to 99.8% confidence intervals)). This allows us to clearly see where a data point falls outside that range (i.e. is an outlier) and investigate further. Control charts require large numbers of observations, so that a robust average can be established for comparison.

'Run' charts require fewer observations than 'Control' charts and (while they may incorporate a median) do not include an established mean value nor control limits to define deviation. While they can be used to measure time between events (often adverse events), run charts are primarily used to measure the number of events over time (in this case, they would feature a series of regular time intervals within the chart). Run charts have specific rules for defining non-random variation over time:

1. A **shift**: six or more consecutive data points either all above or below the median. Points on the median do not count towards or break a shift.
2. A **trend**: five or more consecutive data points that are either all increasing or decreasing in value. If two points are the same value ignore one when counting.
3. **Too many or too few runs**: a run is a consecutive series of data points above or below the median. As for shifts, do not count points on the median: a shift is a sort of run. If there are too many or too few runs (i.e. the median is crossed too many or too few times) that's a sign of non-random variation. You may need to use other resources to establish what an appropriate number of runs would be. An easy way to count the number of runs is to count the number of times the line connecting all the data points crosses the median and add one.
4. An **astronomical data point**: a data point that is clearly different from all others. This relies on judgement. Every data set has a highest and lowest. They won't necessarily be an astronomical data point. Different people looking at the same graph would be expected to recognise the same data point as astronomical (or not).

For further information, see PHI's guidance on Statistical Process Control.

3.3 Charting data to highlight trends

3.3.1 Logarithmic scales

A **logarithmic scale** is non-linear and is based on orders of magnitude; this means that each mark on the scale is the previous mark multiplied by a set value. Logarithmic scales are useful in instances

[4] More information may be found at: http://www.apho.org.uk/resource/view.aspx?RID=39445 [opens Excel spreadsheet], or http://isdscotland.org/Health-Topics/Quality-Indicators/Statistical-Process-Control/

where trends may be difficult to spot, especially where one or more groups is much larger than any of the others (e.g. gender, age group). By using the log of values rather than the values themselves, the data becomes more manageable and easier to interpret. Examples of the use of a logarithmic scale can be seen in Appendix 2.

3.3.2 Changing axis ratios

Another consideration when charting data is the length of the axes. When you create a chart or graph in Excel, it automatically sets the ratio off the axes as 1:1; however this may not be appropriate for your data. An unnecessarily long X-axis can influence interpretation of the data and as a result, an analyst may miss important trends, or the magnitude of those trends. By altering the length of the X-axis (e.g. to half the size of the Y-axis), you may be able to see the scale of a trend more clearly. For an example of this, see Appendix 2.

4. Understanding trend data for exploratory analysis

The analyses described above should be taken into account when deciding whether to further explore data. However, the main consideration is often the customer's requirement. In many cases, a simple description of the trend will suffice. However, if further explanation is required, it may be necessary to investigate the outcome (or dependent) variable in greater detail and to form a hypothesis about what other variables/factors may have influenced observations.

This section describes methods of further exploring data in order to inform decisions about which type of exploratory analysis may be appropriate. The tests described in this section are used to determine which independent variables have a significant correlation with the dependent variable and with the other independent variables. Once these have been identified, it will be perform an analysis to estimate the extent to which the independent variables, individually and collectively, impact the dependent variable.

4.1 Univariate analysis

Charting the frequency distribution of your variables using a histogram (see Appendix 2) may be helpful. The distribution of your data may help to determine what analysis you use; a number of regression analyses (e.g. linear regression), assume that the data is normally distributed and issues may arise where this isn't the case. When analysing a single variable, the most common distribution of data is called a normal distribution. A normal distribution is one where the mean (average), median (middle) and mode (most common) of the data are similar, and the rest of the data are dispersed in a predictable way. When the frequency of values within a variable are plotted as a histogram, if normally distributed, a bell curve should be seen. However, there are many types of distribution other than a 'normal' one, many of which are perfectly valid.

Examining the **standard deviation** (see Appendix 2) of your data may also be a factor in deciding what analysis to run. The Standard deviation (SD) is a measure that calculates the amount of variation in an indicator. A high or large SD indicates a high degree of variation from the mean. What is considered a high or low SD depends on the data. However, one method for determining if the SD is large is by dividing the SD by the mean – if the result is greater than or equal to 1 it may be considered a large SD. Assuming that data points from an entire population would be normally distributed, a high SD value may indicate that the data points are overly dispersed and that the sample is not representative

of the population. If this is the case, you may wish to consider the impact of data quality issues or consider modifying the dataset.

Frequencies, standard deviations and comparison of mean and median values are useful for identifying **outliers**. Outliers are values that are markedly different from the range of most observed values and may affect subsequent analyses. Outliers may be genuine measurements, but may also be an artefact of data collection or quality issues. If possible, determine if these are valid data points or the result of such issues and then make a decision on whether to remove these data points. This decision should always be documented in syntax and possibly also in statistical outputs. The parameters used to identify and exclude outliers may vary on the basis of the number and distribution of values, and can be quite subjective. There are many tests that may be used to identify outliers, such as Tukey's test.[5] If you decide not to exclude outliers, be aware that they may skew analysis, although some of the methods described in the guidance (e.g. smoothing) may help avoid this.

4.2 Multivariate analysis

Having investigated the characteristics of the outcome (dependent) variable, it is appropriate to start to form hypotheses about what other variables (predictors or independent variables) may be influencing observations. Start with the most plausible explanations based on your knowledge of the subject area. For instance, the most plausible explanations for an increase in long A&E waits would be an increase in patients presenting at the A&E or a decrease in staff numbers. In many cases, you will be in possession of the relevant data to compare with the outcome variable. But in other cases, it may be necessary to source this from elsewhere.

Crosstabulating or charting two variables together (for instance, using a scatterplot) is a useful way of assessing the nature and strength of the relationship between them. If there is no relationship, we would often reject the hypothesis that there had been a simple interaction between these variables. However, if we observe a potential association, then it may be appropriate to investigate further and attempt to quantify the strength and nature of the relationship and to ascertain how this interaction occurs.

Some of the methods for doing this are described in detail below. When choosing which method to use, try to adhere to the following rules.

- ✓ ***Keep it simple:*** choose a technique that does the job but isn't more complicated than necessary. This will cause less confusion among users of the data.
- ✓ ***Test the method:*** compare the various methods and use one that is consistent with the time series data and the desired output, as described below.
- ✓ ***Review:*** each time a method is to be chosen, review it and if appropriate, change it.

5. Analytical methods for exploratory trend analysis

Having explored the descriptive analyses described above and decided that a more comprehensive analysis is appropriate; you may decide to use regression analysis to describe the relationship between your outcome (dependent) variable and predictor (independent) variables. There are many types of regression that may be used, depending on your data and the preferred output.

[5] Other tests are described on the Wikipedia page on 'outliers'

It is important to adhere to the principle of parsimony - **the best method to use is the simplest one for the task**. If a chi-square test explains the data adequately and provides the required output, it's unnecessary and unproductive to run a complex regression model.

An example of a PHI publication which uses exploratory trend analysis is:

- Hospital Standardised Mortality Ratios (HSMR) – this publication uses linear regression to present a visual impression of trends in HSMRs over time. Linear regression is used to help overcome the seasonal variations observed in the data; this is a good example of keeping the analysis as simple as possible while producing outputs that are relevant and informative.

Regression analyses are used to learn about the relationship between a set of indicators and an outcome. More specifically, regression analysis is used to aid understanding of **how the mean**[6] **of the dependent (outcome) variable changes when any of the independent (predictor) variables is varied**. It is more powerful than a simple correlational analysis[7] since it allows us to estimate the extent to which the outcome variable will change when the predictor variable does. For example, when we look at the likelihood of mortality due to smoking over a number of years, regression analysis will allow us to estimate the actual impact of age, gender or deprivation. Correlational analyses only allow us to measure whether a statistically significant correlation between variables exists.

The choice of regression method depends primarily on three things: the nature of the relationship, the number of predictor variables and the desired outcome.

1. The relationship between variables can be examined by plotting the data points on a scatterplot (see Appendix 2). Attaching a line of best fit to the scatterplot also aids in determining the nature of the relationship; most statistical packages (including SPSS and R) perform this function.
2. The number of predictor variables to be used in the model is important; one independent variable requires the use of univariate regression (e.g. simple linear regression) while multiple or multivariate regression (e.g. multiple linear regression) is used when two or more independent variables are to be included.
3. Regression analysis may be used to determine if there is a significant difference across years in a time series or it may be used to predict events in the future. It may also be used to identify the impact of a change, for example in policy. The type of regression used depends very much on which of these outcomes is desired.

5.1 Using Regression

In regression, the ideal model is one which is comprehensive enough to fit the data well but also straightforward enough that it interprets and smoothes the data in a useful way instead of 'overfitting' it. Once the model has been created, there are a number of ways to check the model's efficacy:

✓ **Goodness of fit**: these tests describe how well the constructed model fits the observations that have been included. It does this by examining the difference between the observed values and those expected under the model. The most common way of doing this with regression analyses is by using the coefficient of determination, or R(r)-squared.

[6] Mean is used most often. However, in some cases other location parameters (e.g. quantile) may be used.
[7] A correlational analysis is an analysis performed to find the strength of the relationship between two or more variables, e.g. the relationship between height and weight.

- ✓ **Analysis of patterns of residuals**: residuals are the differences between the predicted and observed values in a dataset. The sum of the residuals within a random sample must be close to or equal to zero. A suitable test of residuals is the White test.
- ✓ **Statistical significance of model**:
    - The overall fit can be checked by an **F-test**. This test compares the ratio of two variances to test if they are equal. It uses the assumption that the null hypothesis is true (that interactions between dependent and independent variables are a result of sampling or experimental error and not due to the effect of one on the other). The F test statistic is produced in the output of regression analyses in SPSS, and is described in Appendix 2. A large F-value[8], especially paired with a very small p-value[9], indicates that variation is unlikely to be due to chance. A large F-value indicates that a model is likely to be robust and will produce meaningful results.
    - The statistical significance of the combination of predictor variables is examined using **t-tests**. This test is used to determine if two groups (e.g. individuals aged under 40 and individuals aged over 40) are significantly different from each other in relation to an observation (e.g. alcohol consumption). It assumes that the data are normally distributed. If the t-test shows that alcohol consumption is statistically significantly different for younger and older individuals, it may be assumed that the variables alcohol use and age will be of use in the regression model.

Caution should be taken when interpreting the results of regression, as it is still a correlational method and, as such, causality cannot be inferred from the outcomes.

5.2 Methods

5.2.1 Linear regression

This is recommended for use where the dependent data are normally distributed (e.g. blood pressure) and the independent variables must also have a linear relationship with each other (e.g. weight and height). Linear regression can be done with only one independent variable (i.e. simple linear regression) or with multiple independent variables (i.e. multiple linear regression).

Before performing linear regression, it is important to understand the assumptions that must be met for it to be a robust method of analysis[10].

- The dependent variable should be measured on a **continuous scale** (e.g. years). The independent variables may be continuous or categorical (e.g. age groups), though adjustments may be necessary in the case of categorical variables.
- **The independent variables may not be a combination of any of the other independent variables**, e.g. an age/sex grouping must not be used alongside age or sex.
- **Independent variables must be uncorrelated** with each other. This is considered true if the presence of one does not affect the probability distribution of the other (e.g. height is uncorrelated to blood pressure; height is correlated to weight).

[8] The further away from 1 the F-value is, the less likely that the differences are due to chance. More information on calculating it can be found at Statistics How To (2013).
[9] The p-value is a measure of the significance of observed differences between two groups in a statistical model.
[10] More information on linear regression assumptions can be found at Statistics Solutions.

- The data must show homoscedasticity, or **homogeneity of variance**. This means that the variation within each of the populations must be the same.
- The **variance** between observed and predicted values must be approximately **normally distributed**.

Linear regression can be a powerful analysis if the relationship between the variables is close to perfectly linear[11]. The less linear it is, the less robust the output will be. Additionally, since linear regression compares the mean of the dependent variable to the independent variables, it may be of limited value when examining atypical outcomes (for example, if the aim is to measure low birth weight against other factors, linear regression may not be useful). It is important to remember that it is not an absolute description of the relationship between the outcome variable and the predictor variables. An example of linear regression performed in SPSS is shown in Appendix 2. The linearity of a relationship may be tested using restricted cubic splines, as explained below.

5.2.2 Logistic regression

Of the methods described, linear and logistic regressions are the most frequently used. The key difference between the two is that linear regression determines the relationship between the dependent and independent variables where the dependent variables is measured on a continuous scale (e.g. years), while logistic regression uses a dependent variable with a limited number of potential outcomes (e.g. yes/no/maybe; 1/2/3/4; died/survived).

Logistic regression is most often used when the outcome is binary (e.g. dead or alive after a certain time period). Logistic regression calculates the probability of the binary outcome (i.e. dependent variable) given various values of the independent variable. It can be used for dependent variables with more than 2 outcomes, but the values assigned will be arbitrary (e.g. age groups) and have no relationship to the count of observations.

Logistic regression assumes that the **observations are independent**, e.g. one person dying after a specific diagnosis does not affect the probability of death of another person with the same diagnosis. It is also assumed that the natural logarithm of the odds ratio[12] has a **linear relationship**[13] with the independent variables.

The dependent variables and residuals don't have to be normally distributed for logistic regression to be of use. A linear relationship between the dependent and independent variables is not assumed. However, a very large sample size is needed for robust results.

See Appendix 2 for an example of how to perform logistic regression in SPSS.

5.2.3 Polynomial regression

This method may be appropriate if a scatterplot of the data shows one or more curves. The most common way of interpreting this type of data in regression is to fit curves to the data using polynomial

[11] A linear relationship can be seen clearly on a graph; when plotted with one variable on each axis the points will join to form a line. This relationship can be positive (both variables increase at a steady rate) or negative (one variable increases while the other decreases at a constant rate).

[12] Odds ratio is the ratio of the likelihood of an event being observed in one group to the likelihood of it being observed in another group. An odds ratio of 1 means that the event is equally likely to be observed in both groups; greater than 1 suggests the event is more likely in the first group; less than 1 suggests the event is more likely in the second group.

[13] This can be checked with a Box-Tidwell test. A demonstration of how to do this in SPSS is shown on page 10 of http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.pdf [accessed 3 March 2017]

terms (e.g. squared or cubed independent variables). The polynomial term chosen is usually determined by how many curves are needed[14] to fit the data, but rarely exceeds a cubed term. Statistically this is considered linear regression because the parameters are linear (e.g. years).

In polynomial regression, as in linear regression, a number of assumptions about the data are made.

- The dependent and any or all independent variables have a **linear or curvilinear relationship**.
- The **independent variables are also independent of each other**.
- The **errors are independent (i.e. no correlation between consecutive errors) and normally distributed**.

It is a more complicated method than simple linear regression, and lines of best fit curve oddly and implausibly at their extremes. It can also be more difficult to interpret the output. See Appendix 2 for an example of how this method is done.

5.2.4 Restricted cubic splines[15]

This is a method of testing the linearity of relationships between variables (determined by plotting the data on a scatterplot, and interpreting the R-squared[16] values) and summarising non-linear relationships. It may also be used if the non-linearity is of specific interest.

This method splits up the range of values of the independent variable using 'knots' to define the start and end of each section. Each section has a separate curve applied to it. The number of knots is the most important aspect of this model. In general, for small sample sizes (i.e. under 100 cases) three knots should be used. This ensures that there are enough data points between the knots to fit each polynomial. Where there is a large sample size (i.e. greater than or equal to 100 cases), more than five knots may be used if the relationship looks like it changes quickly and often. Studies (e.g. Harrell et al., 2001; Stone, 1986) have suggested that 4 or 5 knots may be sufficient for the majority of instances where the restricted cubic splines method is employed.

The location of the knots may be determined by the analyst but it is more common for the knots to be pre-specified based on the quantiles of the continuous dependent variable. A table with suggested locations expressed in quantiles of the continuous variable is found in Appendix 2.

Appendix 2 also contains a step by step guide to producing a restricted cubic splines model. Once the model has been run, the variables created can then be used in any regression model as the independent variables (including the original variable).

This method is unusual in that you don't necessarily need to consider the desired outcome when choosing this method, as it transforms the data to suit the model you would like to use. It is also a useful tool to formally investigate the linearity of the relationship between variables. In addition, it is less prone to odd curves at the extremes than polynomial regression. However, it's relatively complicated to perform and interpret, especially when compared to simple linear regression.

[14] Quadratic models are used for one curve; cubed models for two curves – see Appendix 2.
[15] This method is called restricted cubic splines because a. the spline (a regression model whose function/slope changes) is restricted to be linear before the first knot and after the last knot; and b. cubic is the smallest degree of polynomial that provides suitable flexibility for fitting data, so is most often used.
[16] The R-squared (R^2) value is a measure of how well the data are to the line of best fit. Usually, the closer to 1 that R^2 is, the better the data fits the line. However it should be used in conjunction with other test statistics and residual plots (see Appendix 2) when determining if the model is suitable.

5.2.5 Segmented regression

Segmented regression is useful for interpreting trends before and after a point or points of interest. For instance, it may be used when we want to see the effects of a significant event or change in policy (e.g. smoking prevalence after the smoking ban was introduced). It may also be used in instances where there are missing data for certain time periods. It fits separate lines showing different relationships between variables before and after the point of interest.

Assumptions made about the data when segmented regression is performed are:

- **Linearity** between the variables is assumed, which may not always be true. It may also be that the nature of the relationship may be linear in one section but not in another.
- A **minimum of 8 data points** are also recommended before and after the point of interest.

If the breakpoint (the point at which the relationship between the data changes) is unknown, the sections may be determined in a number of ways.

- The simplest method is to plot the data on a scatterplot and identify where there are differences between one section and another.
- If the breakpoints are not obvious from the scatterplot, regression splines may be used (see Appendix 2).
- Another method for determining where knots should be is Joinpoint Regression. This is usually done in a specific programme called Joinpoint. This programme will calculate where differences between years are significant, up to a maximum defined by the user. Obviously the specificity of the programme makes it slightly restrictive as a possible method.

Once the breakpoints are known, a regression model is fitted to each section, and interpreted as in Appendix 2. The strengths of segmented regression lie in its ability to assess the immediate impact of what changed the relationship between the variables as well as the impact over time. It also controls for changes that would have happened without the intervention (e.g. smoking ban). Interpreting the results can be quite straightforward when the data are graphed; visually demonstrating the effects of policy change and the strength of those effects. It is also more robust than simple pre-post analysis as it highlights trends that may have been in place before the change in policy.

5.2.6 Poisson regression

This method is used when the outcome is a count of observations (e.g. hospital admissions) and a Poisson distribution is assumed. A Poisson distribution is the probability distribution that results from a Poisson experiment. A Poisson experiment is one where:

1. outcomes are events occurring or not occurring;
2. the average number of events is known;
3. the probability of the event happening is proportional to the amount of time measured; and
4. the probability of the event in a small amount of time is almost 0.

An example of this may be determining the probability of having 200 people at A&E on Tuesday, if we know the average number of people who are seen at A&E on a Tuesday is 180.

A Poisson random variable is the number of events seen in a Poisson experiment. The probability distribution of a Poisson random variable is called a Poisson distribution. This has specific properties:

1. the mean of the distribution is the mean of the number of observations.
2. the variance is also equal to the mean of the number of observations.

As with all types of regression, several assumptions are made about the data.

- The events must be **independent of each other**, i.e. the occurrence of one observation will not make another more or less likely.
- The number of events should follow a **Poisson distribution.**
- Independent variables must be continuous, dichotomous or ordinal.
- Events must be counted in **positive whole numbers**.

Poisson regression is especially useful for estimating rare events, such as calculating the probability of a patient being diagnosed with a rare (non-infectious) illness. It is similar to logistic regression but a robust model can be obtained with Poisson regression using much lower sample sizes.

As mentioned previously, Poisson regression may also be used to determine if changes occurring across a time series are significant, while adjusting for an independent variable, such as age. Before proceeding with this analysis, it must be decided if the trend is reasonably linear or if another type of relationship should be explored. Similarly, seasonal patterns should be considered, as this may account for much of the change seen.

An example of a Poisson regression analysis in SPSS can be found in Appendix 2, with guidance on how to interpret the outputs.

5.2.7 ARIMA

ARIMA (AutoRegressive Integrated Moving Average) models take account of the correlation between successive data points. It requires the data points to be relatively stable over time and with a minimum of outliers. It is also recommended to have at least 40 data points when using ARIMA models. ARIMA modelling is mainly used for short term forecasting.

Before using an ARIMA model, the data should be checked for 'stationarity' - a measure of the predictability of changes over time. Decreases in mortality every June is not stationary (it may be predicted), while random variations observed within a year may be stationary (not predictable). Stationarity is required for ARIMA modelling. However, if it is not present within the data, it is possible to 'difference' the data to transform it to stationary data. This is done by subtracting the observation in the current time period, e.g. month, from the observation in the previous time period ('first differencing'). If this is not sufficient to make the data stationary, this process may be carried out again – referred to as 'second differencing'. The purpose of differencing is to stabilise the mean.

Other assumptions of ARIMA are:

- all observed time series' are random
- there are no other predictors
- the relations are exclusively linear

ARIMA models can be autoregressive, or have a moving average, or a combination of the two. Autoregressive models forecast the dependent variable based on a weighted sum of past values. The independent variables in this instance are lagged values of the dependent variable.

Moving average models use past forecast errors in a regression-like model. These models use a value that is not observable, which means that calculating it presents some issues, and iterative non-linear fitting procedures need to be used.

SPSS has a function that can decide which ARIMA model is best for your data (see Appendix 2). It may include independent variables (specified by you) but only ones that have a significant relationship with

the dependent variable. It also differences or uses other methods to make the data stationary. There is another function that allows you to build the model yourself but this should obviously only be used once you have a lot of experience using ARIMA models.

5.2.8 Age Period Cohort Analysis

This method is used to form projections of incidence and mortality (e.g. in relation to medical conditions), based on three effects within a population: age (e.g. varying rates of disease), time period (e.g. social or economic factors) and cohort (e.g. difference in health outcomes based on year of birth). This method will be described in further detail in forthcoming PHI guidance on predictive analysis.

See Appendix 1 for an example of age period cohort analysis from PHI's Cancer Incidence Projections.

6. Discussion

There are a wide variety of methods for producing meaningful trend analyses, many of which are described in this guidance.

While it is valuable to learn about options for analysing trend data and it may be tempting to apply these techniques within data analysis, it is more important that analysts:

- understand their data;
- analyse descriptively first; and
- if undertaking more complex analysis, select an appropriate method that explains data in the simplest manner possible.

Understanding the ways in which data are collected, submitted, validated and coded will also help with your analysis. For instance, you may have missing values that will need to be considered in the context of any analysis. Issues found in the process of understanding the data may explain observable differences; if they don't, further investigation may be required.

Once you have determined that your data is of sufficient quality to be analysed, first analyse descriptively. In many cases the customer will not require anything more detailed. For example, running a complex analysis on data that can be explained by seasonal effects is inefficient and is unlikely to add more value than a simple crosstabulation.

While extremely valuable in some circumstances, complex analyses such as regression are time-consuming and require statistical/technical knowledge. Where a customer does wish to explore the data further, ensure all stakeholders (internal and external) are aware of the impact of that decision on timescale and resource. Ensure the specification/timescale allows the analysts to proceed through the steps outlined above (descriptive analysis, hypothesis building, regression analysis, testing) without undue pressure from stakeholders. Communicate findings to internal and external stakeholders on a regular basis in order to validate decisions and test assumptions.

Appendix 1: Examples from PHI outputs

1.1: Simple Smoothing and Confidence Intervals

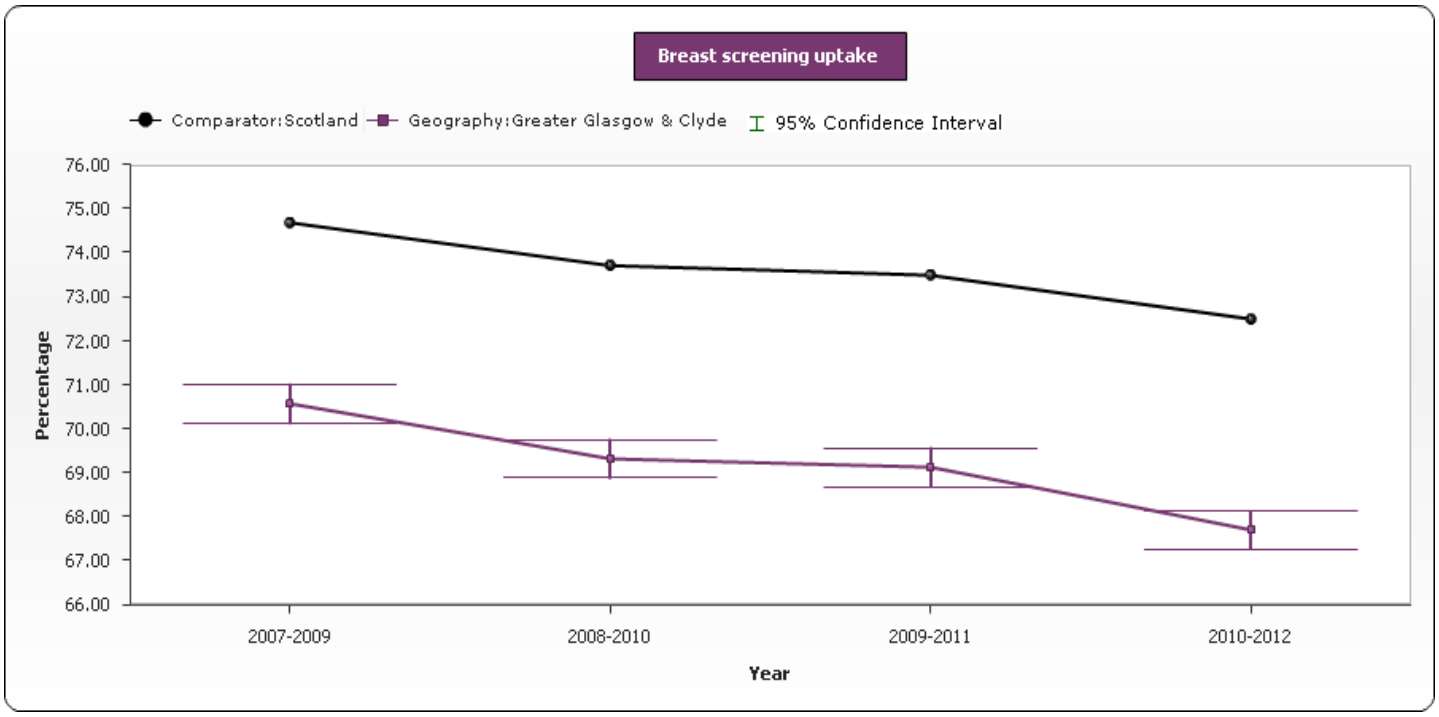
ScotPHO Health and Wellbeing Profiles: Breast Screening Uptake (NHS Greater Glasgow and Clyde)[17]

The ScotPHO Health and Wellbeing Profiles provide a snapshot overview of health for each area using spine charts which allow comparison to the Scotland average. Rank charts and trend charts (as below) are also included in the Profiles to allow further understanding of the results.

Many of the indicators reported in the Health and Wellbeing Profiles on ScotPHO are reported by local authority, and several report very small numbers of observations. In order to improve the reliability of the data, and to maintain confidentiality, the numbers and rates for each geography level are presented as three-year averages.

Figure 1.1 below shows the three-year rolling averages with 95% confidence intervals for breast screening uptake in NHS Greater Glasgow and Clyde, compared to the three-year rolling averages across Scotland.

Figure 1.1



[17] https://scotpho.nhsnss.scot.nhs.uk/scotpho/homeAction.do

1.2: Linear Regression

Hospital Standardised Mortality Ratios (HSMR)[18]

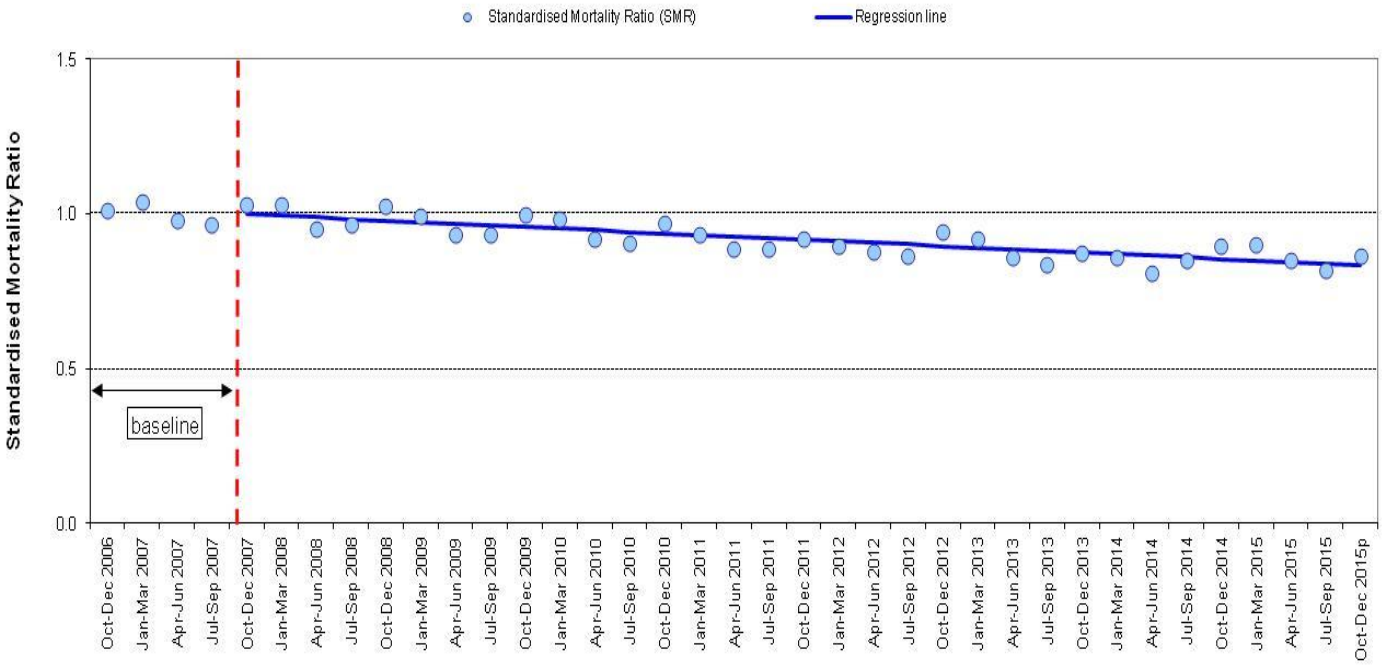
Hospital Standardised Mortality Rates (HSMRs) are produced PHI and reviewed jointly between PHI and Health Improvement Scotland to identify potential patterns in the data. The Scottish Government uses these statistics to monitor change in hospital mortality over and inform policy decision making. NHS Boards use the data to monitor local hospital mortality, and to facilitate improvements in patient care.

HSMRs adjust mortality data to take account of some of the factors known to affect the underlying risk of death. The HSMR calculation is based on patients who died within 30 days of a hospital admission.

A regression line is fitted to the HSMR trend from the first quarter after the baseline period through to the latest HSMR. The percentage change in SMR is calculated by subtracting the regression line value for the baseline from the regression line value for the most recent quarter. The rationale behind this approach is that seasonal variations in HSMR will be smoothed out, and monitoring long term change will be based on a more stable foundation.

Figure 1.2 below shows the standardised mortality ratio for each quarter since October-December 2006, with a regression line based on quarters since October-December 2007.

Figure 1.2



[18] http://isdscotland.org/Health-Topics/Quality-Indicators/HSMR/

1.3: Population Proportion Tests

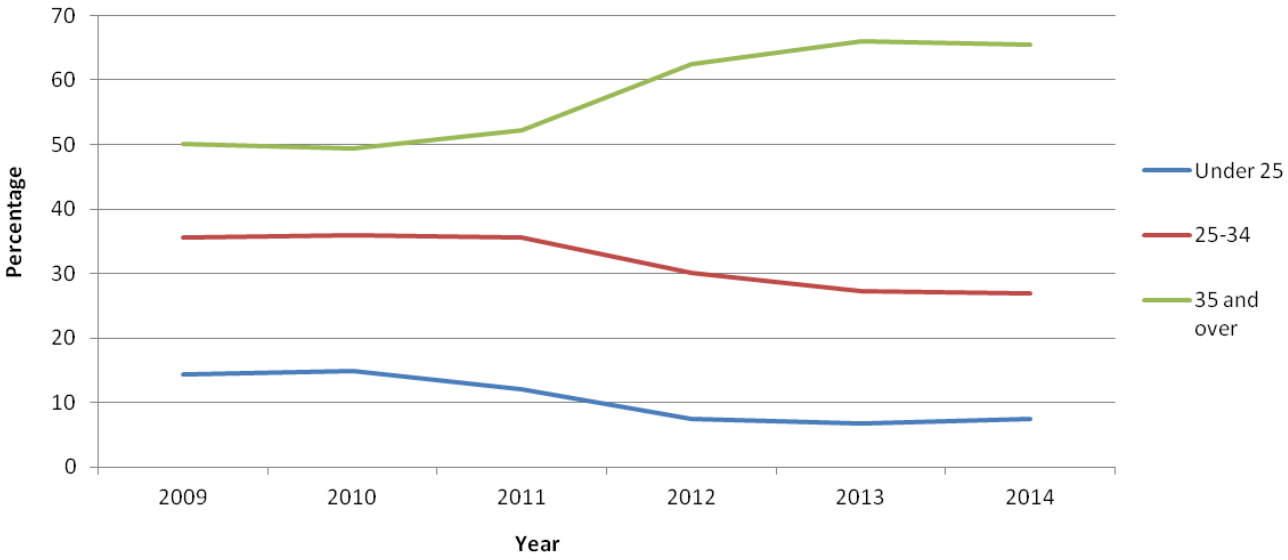
National Drug-Related Deaths Dataset report[19]

PHI Scotland report on the characteristics of individuals whose death was related to the misuse of controlled drugs and the circumstances of those deaths. The rationale for this analysis is to identify behaviours/risks associated with drug-related death and to assess the efficacy of initiatives aimed at addressing behaviours/mitigating risks associated with drug-related deaths.

Within this report, population proportion tests are carried out across the times series to assess the likelihood that observed variations over time are real or the product of differences between annual cohorts of deaths. Wherever a figure has been described as 'higher' or 'lower' than a previous year, the calculated p-value is less than or equal to 0.05.

Figure 1.4 below shows the number of drug related deaths per age group in Scotland for years 2009-2014. Significant differences in the number of deaths among each age group were observed across the time series. These changes are likely to be associated with an increase in the number of older problem drug users.

Figure 1.4



[19] http://www.isdscotland.org/Health-Topics/Drugs-and-Alcohol-Misuse/Publications/2016-03-22/2016-03-22-NDRDD-Report.pdf?

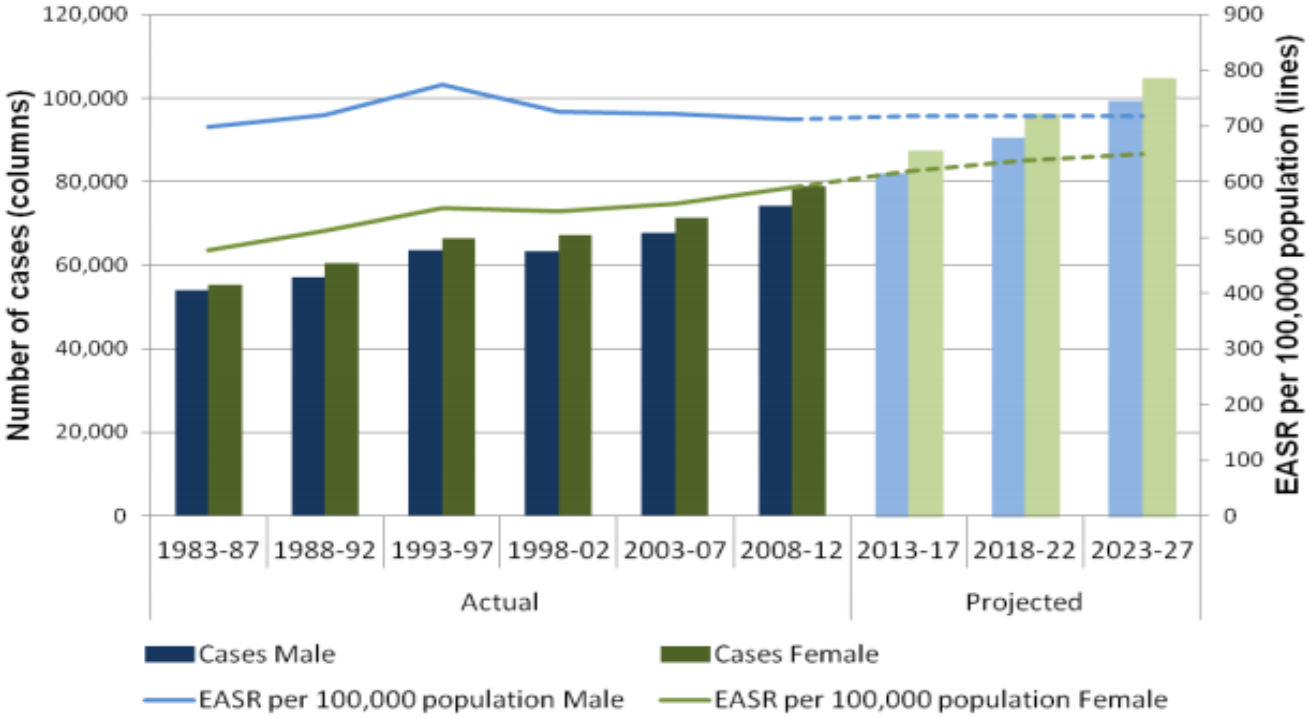
1.4: Age-Period-Cohort Analysis

Cancer Incidence Projections for Scotland: 2013-2027[20]

Every five years, ISD publishes projections for the number of new cancer diagnoses for the following 15 years. The last publication, produced in 2015, predicted the incidence of cancer until 2027, in 5 year increments. These projections were based on Age-Period-Cohort analysis developed by Møller et al (2003)[21]. The analysis assumes that the recent trends in cancer incidence will continue, and that projected populations are accurate.

Figure 1.3 below shows the projected incidence of cancer for years 2013-2017, 2018-2022 and 2023-2027 (with projected European Age/Sex Standardised Rates [EASRs]) alongside past recorded cases and EASRs.

Figure 1.5



[20] http://isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/Incidence-Projections/
[21] https://www.kreftregisteret.no/en/Research/Projects/Nordpred/Nordpred-software/

Appendix 2 Worked examples

2.1: Simple Smoothing

Example of syntax used to calculate rolling averages:

Generate the rolling three year averages figures for the numerator (e.g. N of observations).
Compute AVG_0911 = (n2009 + n2010 + n2011)/3.
Compute AVG_1012 = (n2010 + n2011 + n2012)/3.
Compute AVG_1113 = (n2011 + n2012 + n2013)/3.
Compute AVG_1214 = (n2012 + n2013 + n2014)/3.

The difference between a single-year trend line and the output from smoothing can be seen in the graph below. The example shows Scottish Drug Misuse Database (SDMD) data on the annual number of completed initial assessments for specialist drug treatment for NHS Ayrshire and Arran along with a rolling three-year average. The first point of the smoothed line shows the mean of the total number of initial assessments from 2006-2008; the last point shows the mean of the data from 2012-2014.

Figure 2.1

2.2: Population Proportion Tests

The first step in calculating the significance of the population proportion is to calculate the proportion of events to population (divide the number of events by the population). Next, the Standard Error (SE) must be calculated. Before this can be calculated, p̂ and q̂ must be determined. These are estimates of probable events; p̂ represents the likelihood of it happening and q̂ is the likelihood of it not happening.

To calculate p̂ and q̂:

$$\hat{p} = \frac{N1 + N2}{D1 + D2}$$

Where:

- D1 = denominator 1 (i.e. for first year used);
- N1 = numerator 1 (i.e. for first year);
- D2 = denominator 2 (i.e. for second year); and,
- N1 = numerator 1 (i.e. for second year).

If we use the drug related deaths example, D1 will be total number of drug-related deaths in the first year of interest while N1 could be a count of cases with a given characteristic of interest (e.g. number of people in the most deprived areas of Scotland who died a drug related death in that year).

$$\hat{q} = 1 - \hat{p}$$

Standard error[22] can now be calculated using the following equation:

$$SE = \sqrt{\hat{p} * \hat{q} * \left(\frac{1}{D1} + \frac{1}{D2}\right)}$$

The z statistic is calculated using the formula below:

$$z = \frac{P1 - P2}{SE}$$

Where:

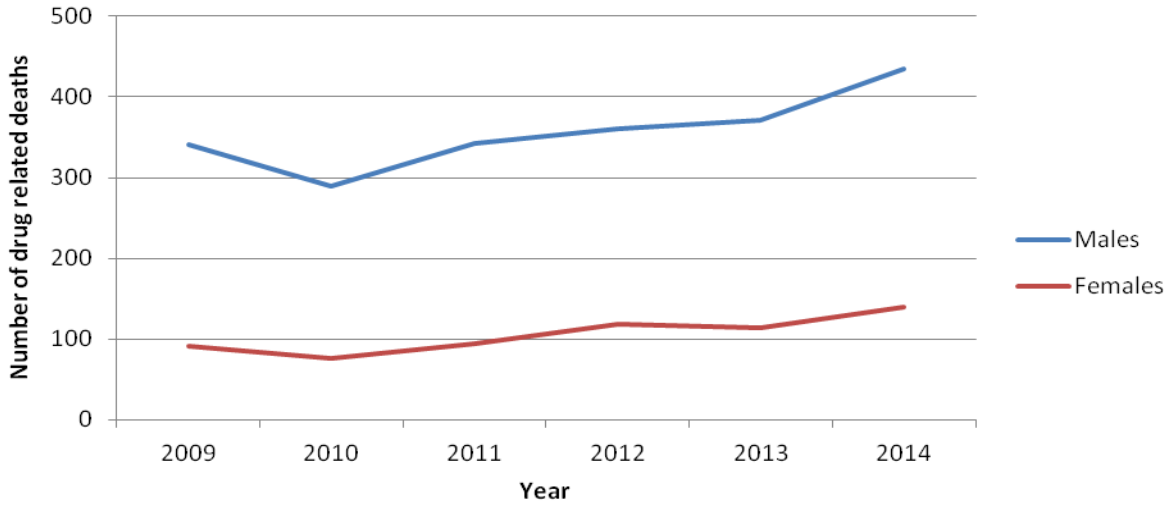
- P1 = proportion 1 (i.e. for first year);
- P2 = proportion 2 (i.e. for second year); and,
- SE = standard error.

Next the p-value is calculated. First, it must be decided if a one-[23] or two-tailed[24] significance test is appropriate. A standard normal cumulative distribution is then calculated (it is not necessary to show this equation as the function "NORMSDIST" in Excel will calculate it for you). If you have decided on a one-tailed test, subtract the standard normal cumulative distribution of z from 1 (i.e. 1-NORMSDIST(z)) and then subtract that figure from 1, (i.e. 1-(1-NORMSDIST(z)). To calculate the two-tailed p-value, multiply the value for a one-tailed test by 2.

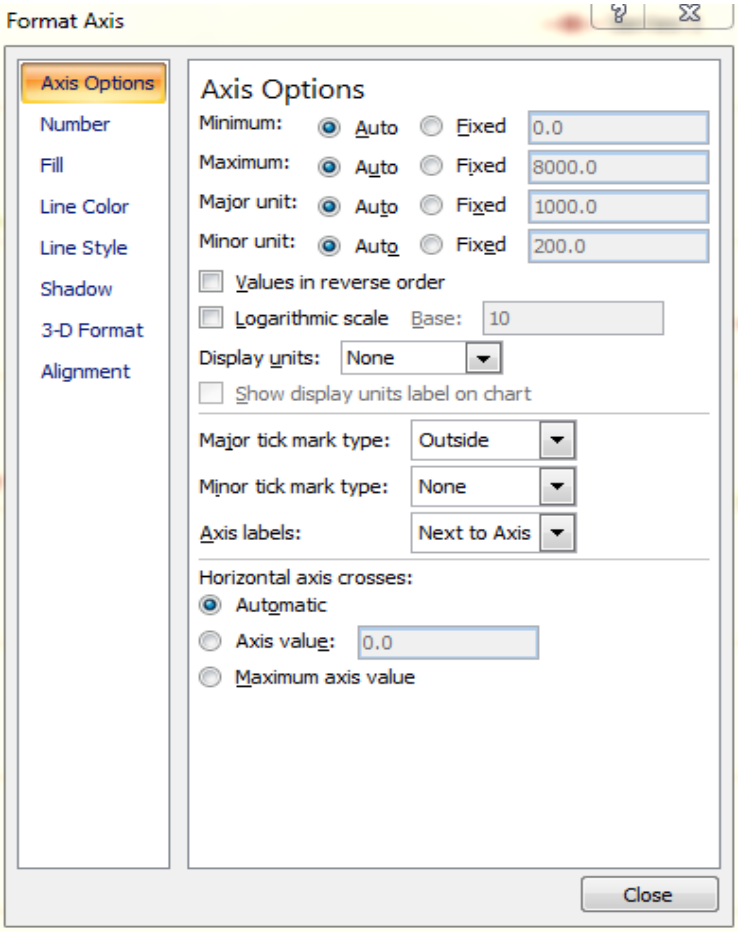
[22] Standard error estimates the variability between the means of the sample.
[23] A one-tailed hypothesis test examines only one side of the data. You can only determine if data are significant in a positive OR negative direction. The other direction is unexamined and no conclusions about it may be made.
[24] If it is assumed that significance is achieved when the p-value is 0.05 or below, a two-tailed test will only be significant if the p-value is 0.025 in either a positive or negative direction. This means that two-tailed tests are more robust than one-tailed tests, but it is harder to achieve a significant result.

2.3: Logarithmic Scales

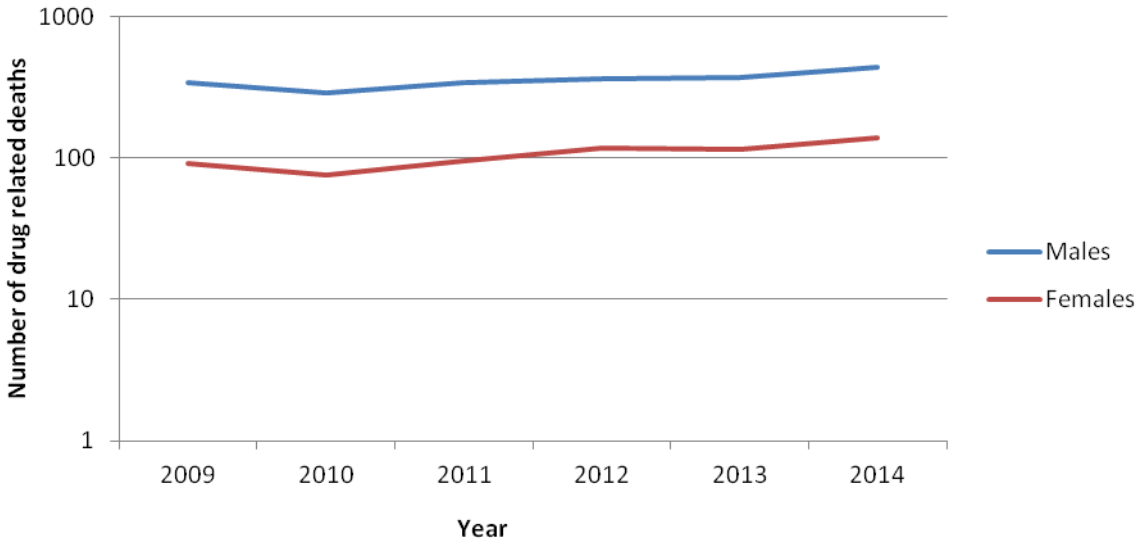
Logarithmic scales can easily be produced when creating charts in Excel. The following chart shows drug-related deaths by sex (2009-2014). The values are plotted on an arithmetic scale.



To change the above to a logarithmic scale, right click the Y-axis, and select “Format axis” from the menu. The following dialogue box will appear:



When the option for ‘logarithmic scale’ is selected, choose the base for the log (the default is 10). This produces the following chart, which provides a different perspective on differences in the numbers of male and female deaths.

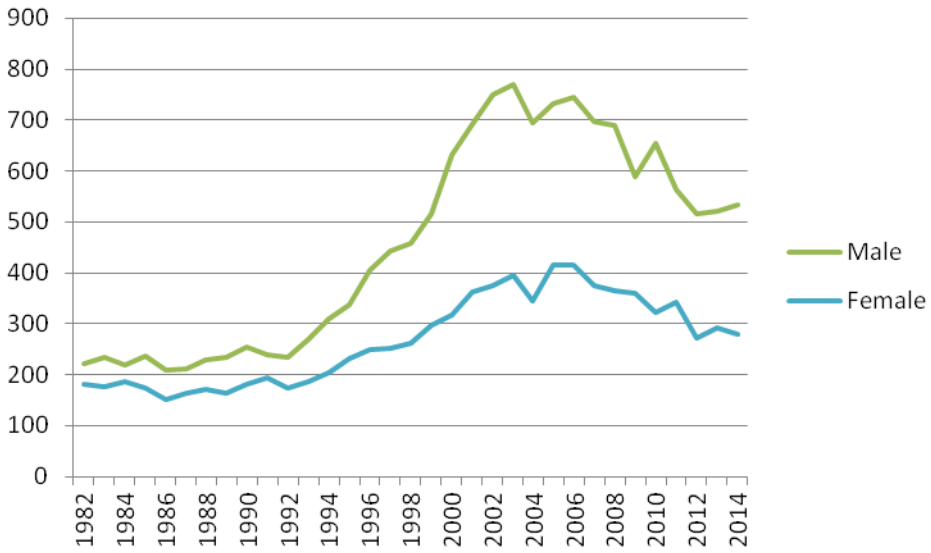


The first chart shows a large increase in male deaths and only a small increase in female deaths. However, the second (logarithmic scale) chart shows that female deaths have increased at a rate comparable to males; this trend was hidden in the original graph because of the large relative size of the male cohort.

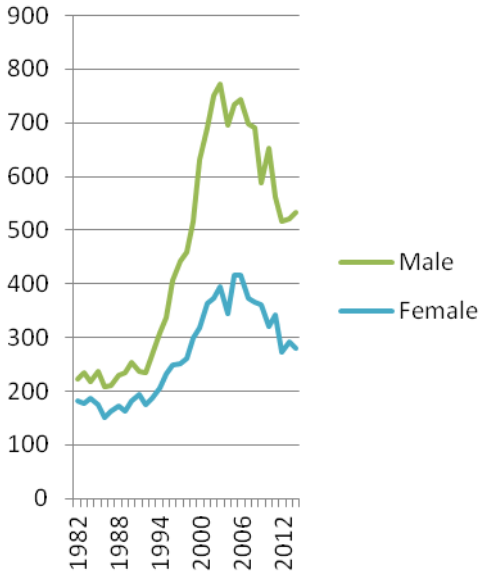
2.4: Modifying axes

Resizing a chart in Excel is a simple matter of clicking the three dots at the extreme left or right, centre top and bottom, or any of the corners of the chart area and dragging in whichever direction you choose. It is possible to resize charts via the menus (Chart tools>Format>Size), but this method resizes the entire chart (including legends etc.) rather than the plot area. For simplicity, resizing the chart by dragging the dots is recommended (the following example resizes the X-axis using the left or right dots).

Below is a graph of number of deaths as a result of chronic liver disease for males and females in Scotland for calendar years 1982-2014, using Excel's generic formatting:



You can clearly see that more males die from chronic liver disease than females and that this is a historic trend. However, when we reduce the width of the X-axis, we see that what looked like a relatively steady change in females is actually more dramatic than the above chart suggests:



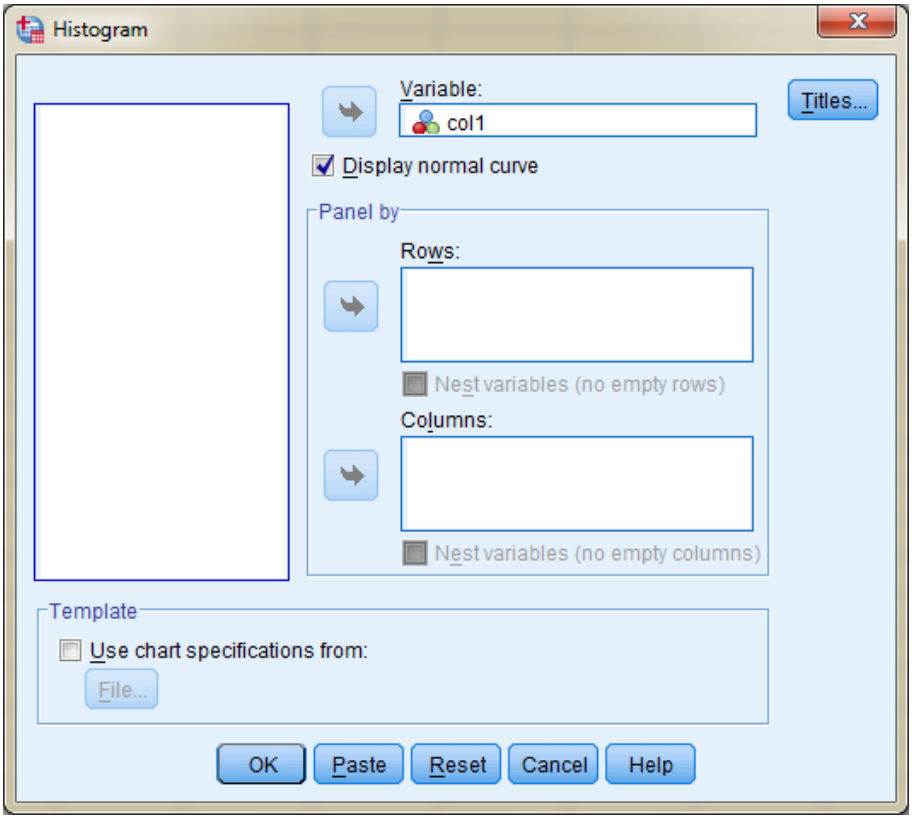
2.5: Histograms

A histogram shows the distribution of a variable (e.g. age group). It differs from a bar chart as bar charts are used to compare categories (e.g. median length of stay in each NHS Board).

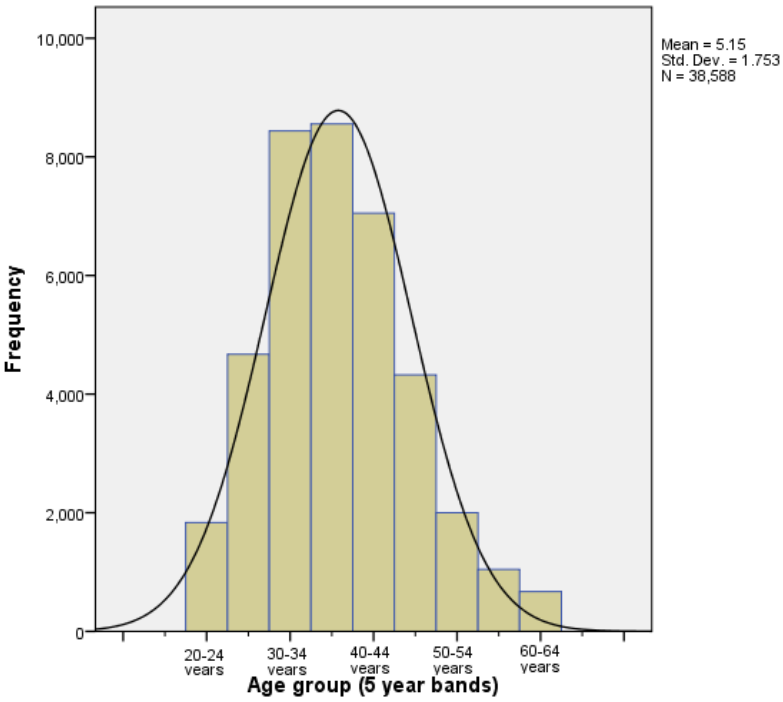
Histograms can be created in SPSS using the following commands:

Graphs → Legacy Dialogs → Histogram

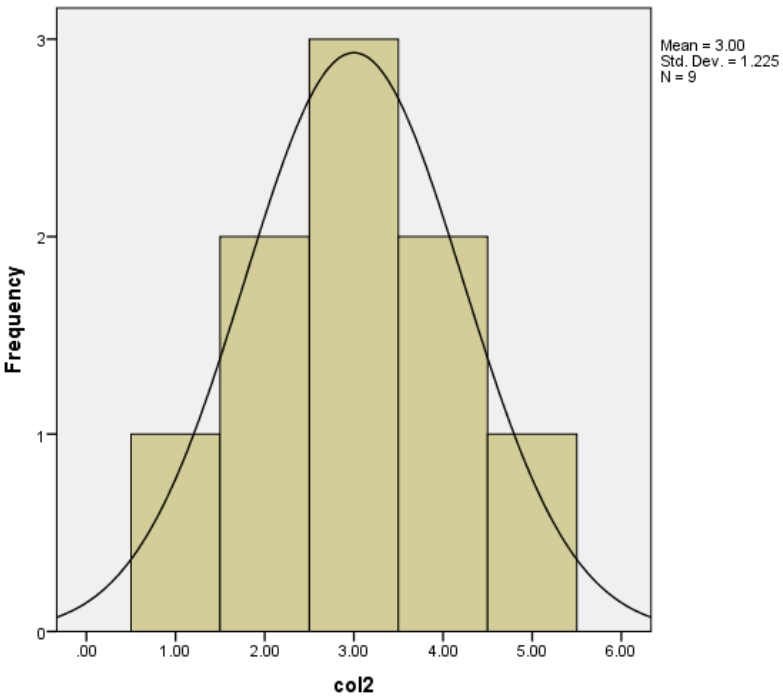
The following menu will be opened:



The following output will be displayed:



The graph above shows the distribution of age in a cohort of problem drug users. We can see that the numbers of people in the extremes of the data (i.e. 20-24 years, and 50-54 years and 60-64 years) are lowest while people from 30-34 years to 40-44 years represent the largest part of the population.



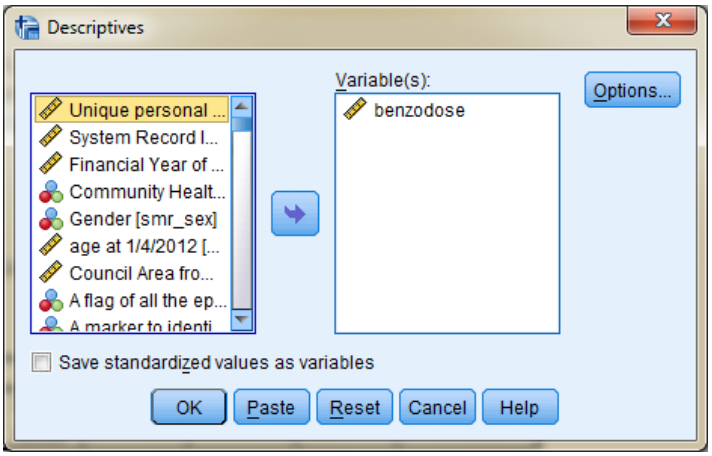
The graph above shows a perfect normal distribution. We can see that the number 3 has occurred most times (3), while numbers 1 and 5 have only occurred once each. This graph was created using dummy data; it is unlikely a true dataset will produce so perfect a normal distribution. However, an approximate normal distribution, as the first graph shows, is sufficient where a method requires it.

2.6: Standard deviation

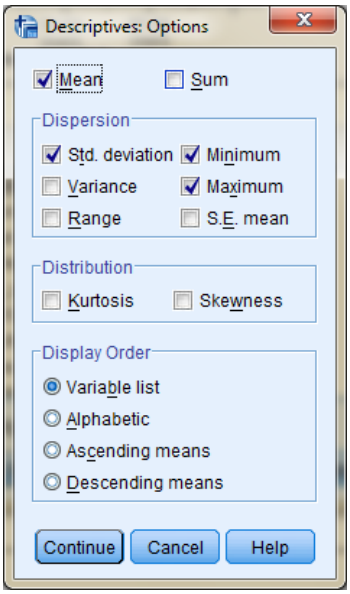
Standard deviation (SD) can be calculated in SPSS using the menus:

Analyze → Descriptive Statistics → Descriptives.

The following dialog box will be displayed:



Once you have selected the variable you would like to examine, click 'Options'. From the following menu, the important options are 'Std. Deviation' and 'mean' (which will help to determine if the standard deviation is considered large).



The resulting table shows us that 1,362 people have a value for 'benzodose', the mean 'benzodose' is 25.89 and the standard deviation is 27.247. Since the SD is larger than the mean, it can be considered a large standard deviation.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
benzodose	1362	0	800	25.89	27.247
Valid N (listwise)	1362				

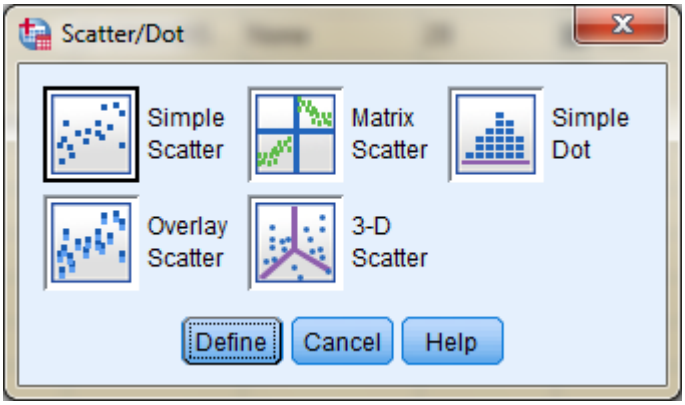
2.7: Scatterplots

Scatterplots can be useful for determining the direction of the relationship between two variables (one dependent and one independent, or two independent). They are also useful for identifying outliers in the data that can then either be controlled for or removed.

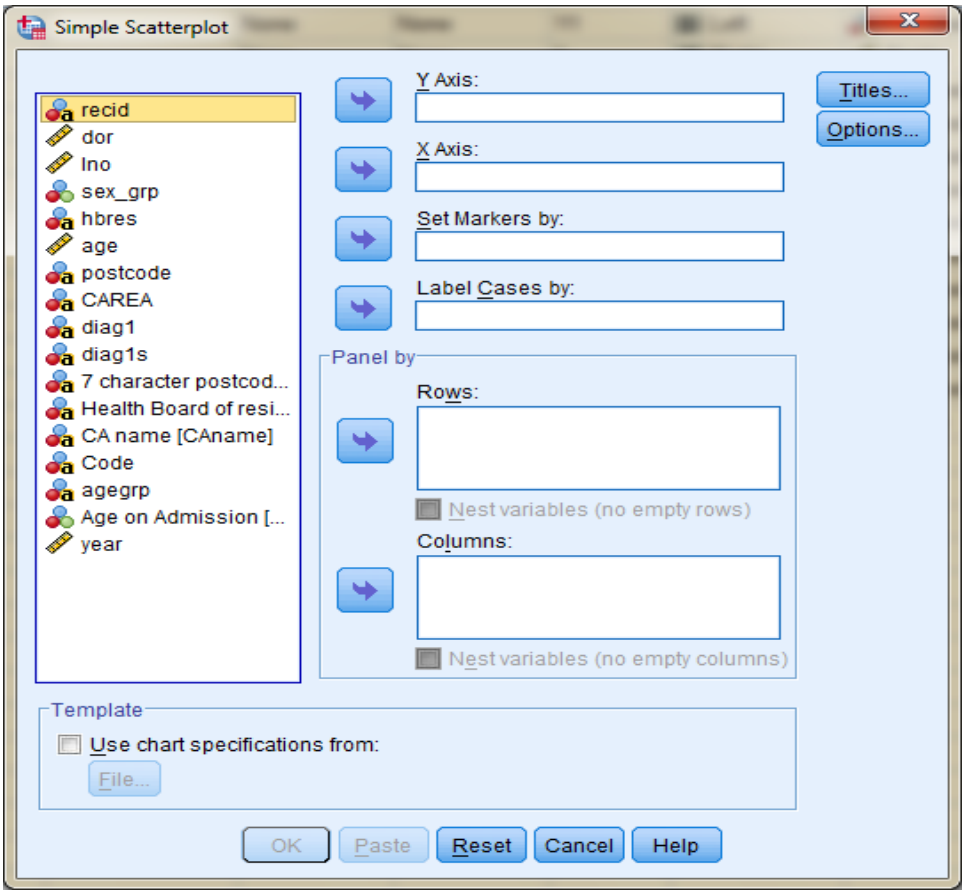
Scatterplots can be created in SPSS using the menus:

Graphs → Legacy Dialogs → Histogram.

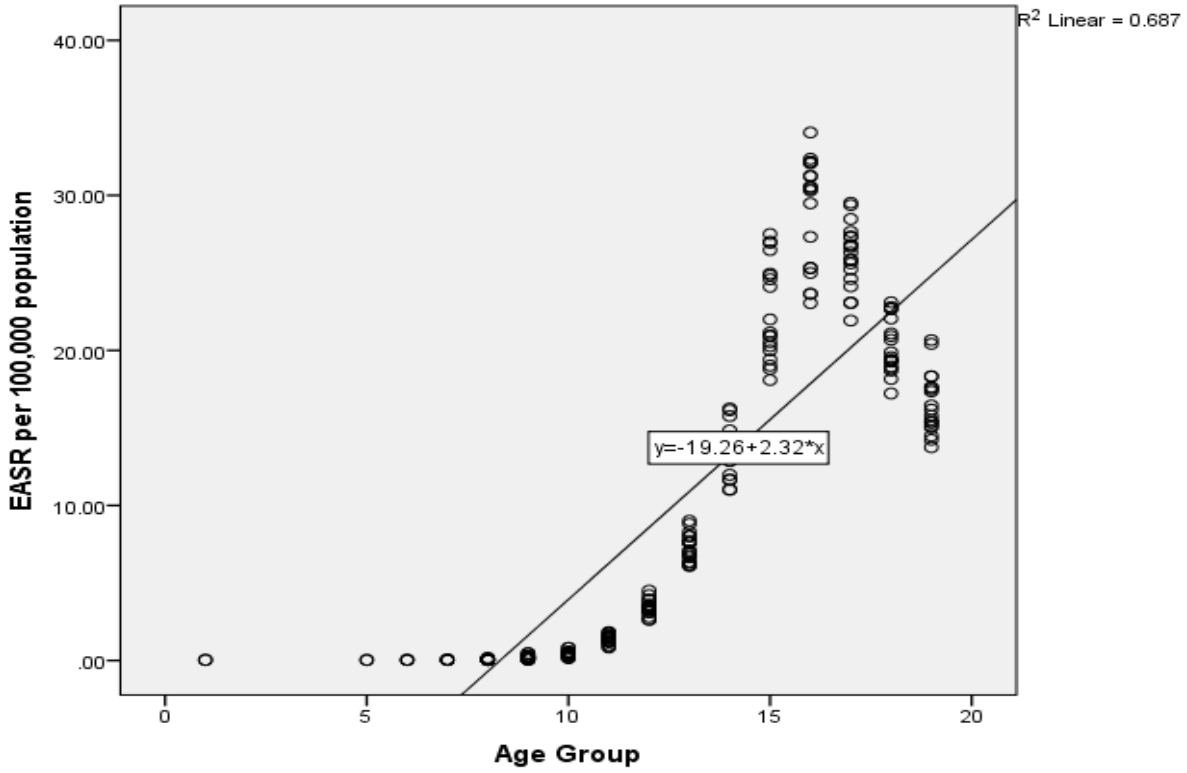
The following dialogue box will then be displayed:



Select 'Simple Scatter' and then 'Define'. Another dialogue box will be displayed (below).



Once you have selected your variables and clicked ‘OK’, a graph similar to the one below should be displayed.



As can be seen, this scatterplot shows a curved relationship between the variables. There are also some outliers which should be investigated before proceeding with analysis.

I have attached a line of best fit to this graph, which draws a line through the data points. The line should be placed at the point where the most data points will be close to it. The equation for the slope of the line is also shown. To fit this line to your scatterplot, double click the graph. Once in the dialogue box, select:

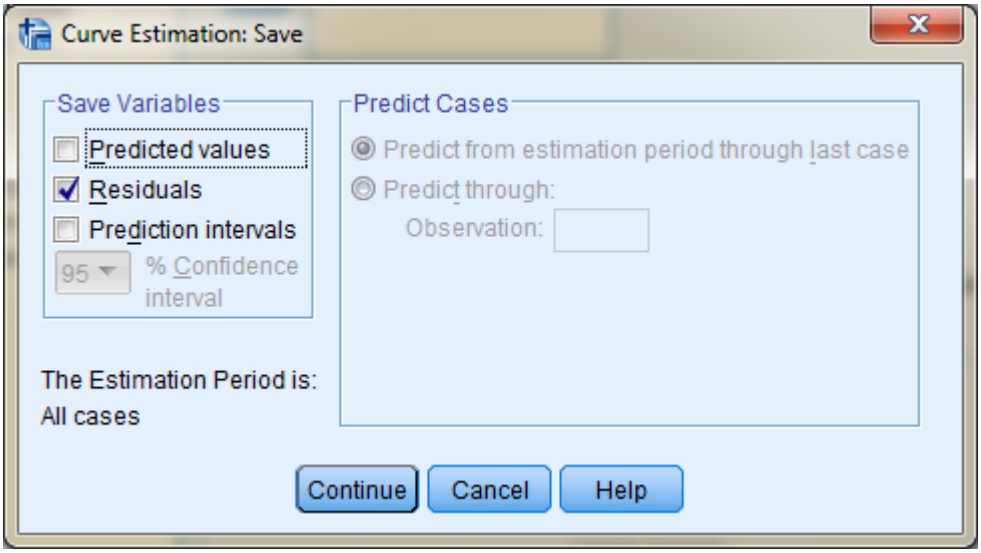
Elements → Fit Line at Total → Linear

2.8: Residual plots

A residual plot is one where the independent variable is graphed on the Y-axis while the residuals, or variations between the predicted and observed values, are plotted on the X-axis. Residual plots are important for determining which form of regression is most suitable (e.g. linear or non-linear (e.g.)).

Residuals plots can be created in SPSS through the menus. Two steps are required:

1. When in the dialogue box for Curve Estimation, press “Save”. Another dialogue box will open, asking which figures you would like to save (see below):



   For each variable saved, a new variable in your dataset will be generated with the relevant data. This is then used in the next step.

2. Create histogram using the newly generated variable(s), as described above

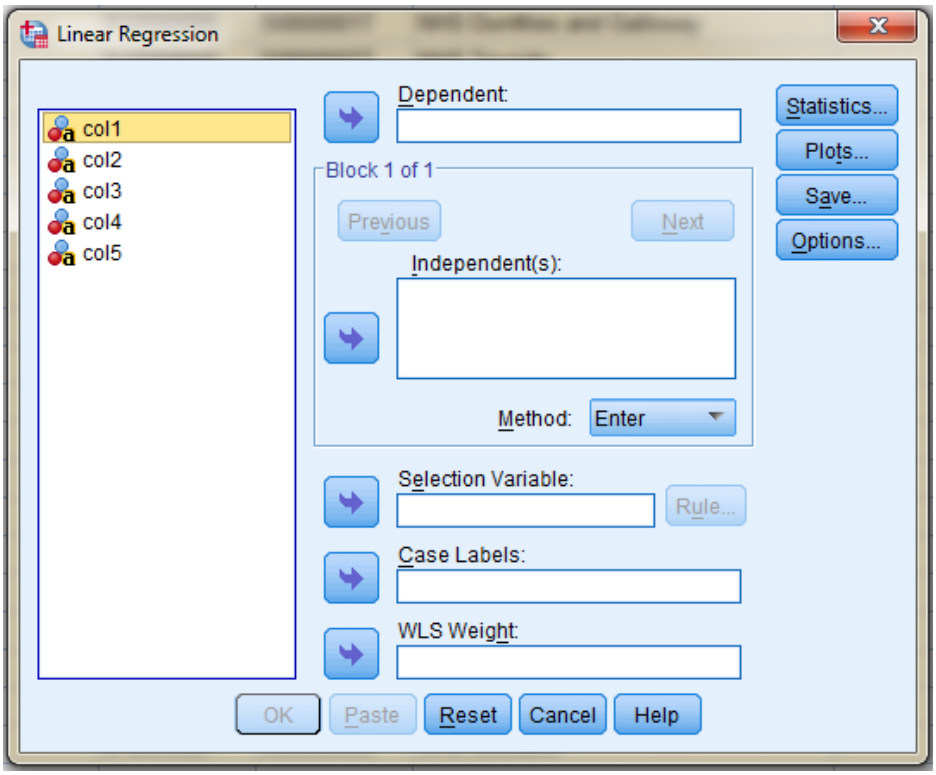
   Once in the dialogue box, ensure that “Display normal curve” is checked. A random dispersion around the horizontal axis indicates that linear regression would be suitable; any other dispersion indicates that a non-linear model may be more appropriate.

2.9: Linear Regression

Performing linear regression may be done manually but as mentioned above, it is possible to do this in SPSS through the menus and R. In SPSS this is done by using:

Analyze → Regression → Linear

The following dialogue box will appear:



Once you have entered at least the dependent and one independent variable, you may press 'OK' and SPSS will run the regression model. The output is a series of tables, which must be interpreted. This example looks at number of deaths from COPD by age group. The first table is 'Model Summary':

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.740[a]	.548	.546	145.853

a. Predictors: (Constant), Age Group

The strength of the correlation is represented by R, which is Pearson's *r*. In this example, it is 0.74 which is moderately strong in a positive direction. Pearson's *r* will always be between -1 and 1. R squared (R^2; 0.55) represents the amount of variance that can be explained by the independent variable (age). R^2 will always be between 0 and 1 (i.e. 0% and 100%). The standard error is an estimate

of the variance of the dependent variable for each age group. This example shows a relatively high standard error[25] which means that predictions made using this model must be treated with caution.

The next table looks at the analysis of variance, or ANOVA.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6001506.090	1	6001506.090	282.118	.000[b]
	Residual	4956620.693	233	21273.050		
	Total	10958126.783	234			

a. Dependent Variable: deaths

b. Predictors: (Constant), Age Group

This table looks at whether the line of best fit is significantly different from 0. The **F-value** is 282.1, and the p-value is less than 0.001, indicating that variations are highly unlikely to be due to sampling error. If the null hypothesis is true, the F-value will be close to 1. The further away from one the F-value is, the less likely that variations are due to chance.

The last table gives the degree to which the independent variable affects the dependent variable. In this example, each age group is estimated to have 42.1 more deaths from COPD than the one before it. The confidence intervals tell us that, when graphed, there is 95% confidence that the population regression line will be between 37.2 and 47.1.

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-329.009	34.455		-9.549	.000	-396.892	-261.126
Age Group	42.124	2.508	.740	16.796	.000	37.183	47.065

a. Dependent Variable: deaths

The constant estimates the likelihood of having a death from COPD when age group (i.e. predictor variable) is 0. In most examples, this figure does not need to be used or reported.

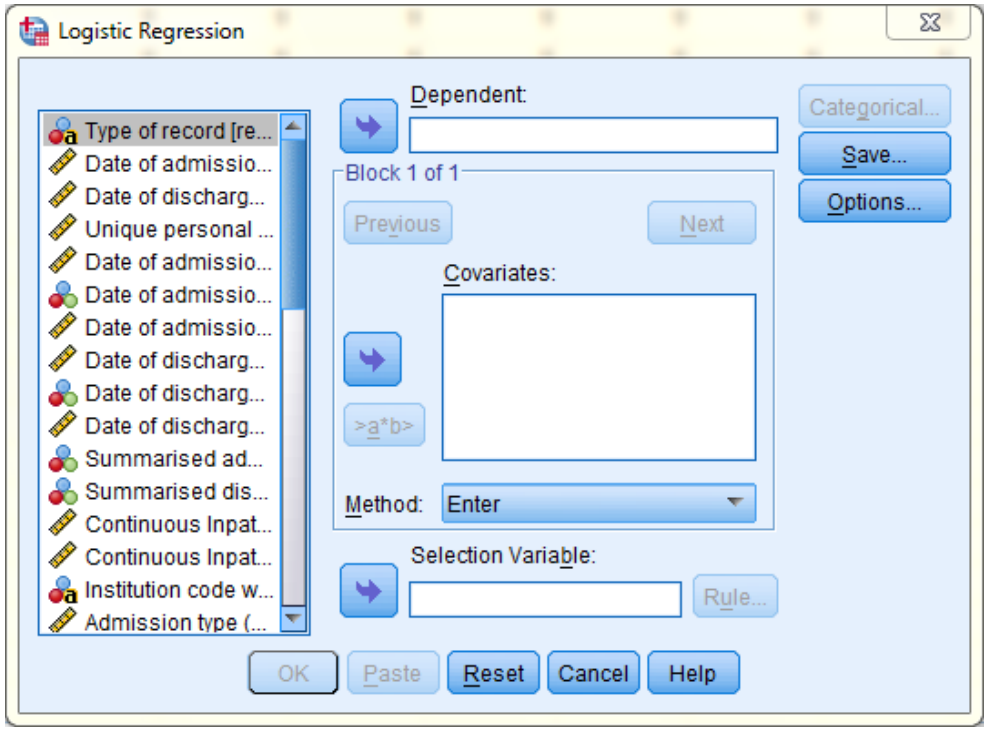
[25] One way of calculating the relationship of the standard error to the mean is to divide the standard deviation by the mean. This is called the "coefficient of variation" (CV); a rule of thumb is that if CV > 1 the standard deviation is large.

2.10: Logistic regression

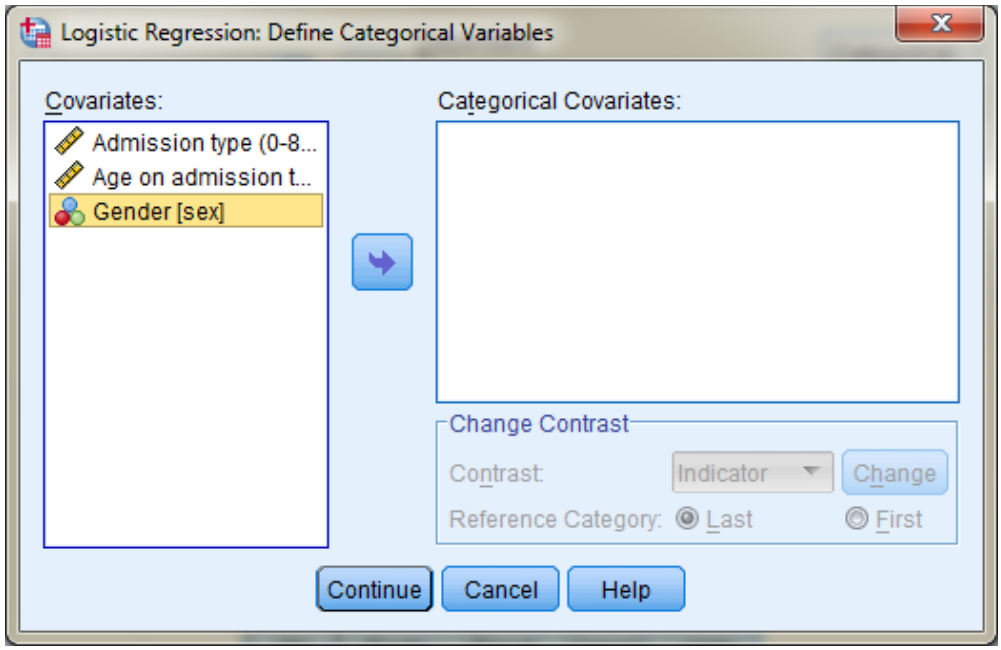
Logistic regression is possible to do in SPSS, through the menus:

Analyze → Regression → Binary Logistic (for a dichotomous variable – two outcomes)
Analyze → Regression → Multinomial Logistic (for more than two outcomes)

This example will look at binary logistic regression. The following dialogue box is displayed:

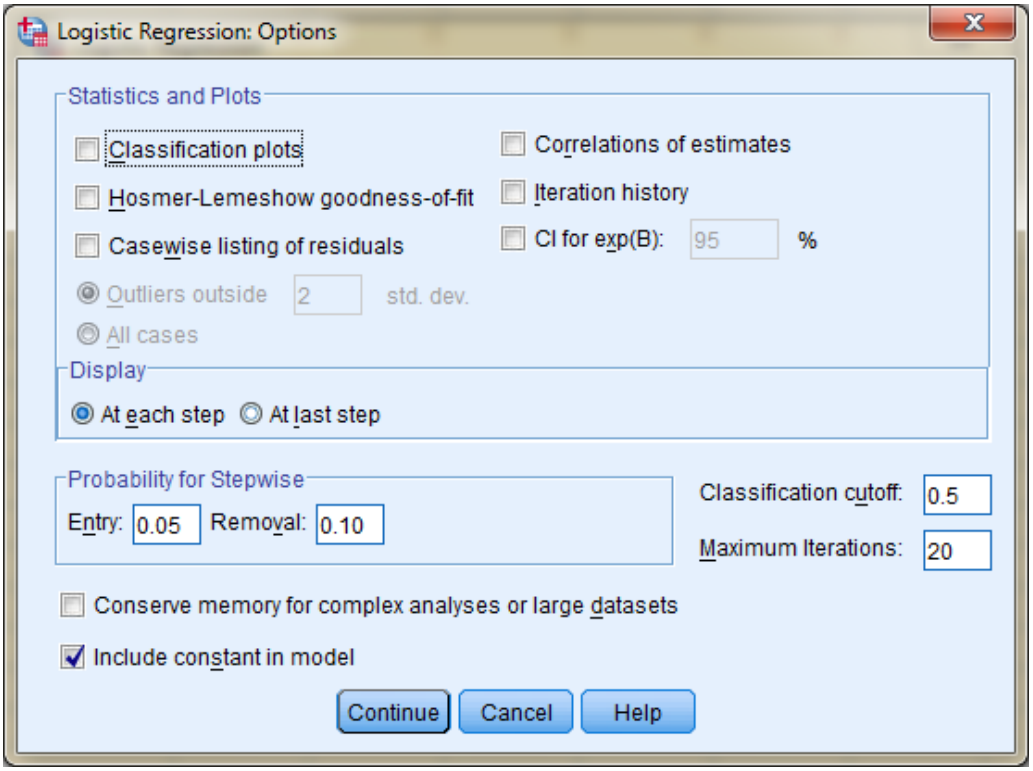


Select your dependent variable and place it in the 'Dependent' box. Your independent variables go into the 'Covariates' box. If any of the independent variables are categorical, choose 'Categorical' for the dialogue box below:



In this example, gender is categorical and must be placed in the 'Categorical Covariates' box as SPSS does not identify these automatically. The reference category is chosen based on what you would like to compare against. If male were coded as 1 and females coded as 2, and you wanted to compare against males (i.e. males is the reference category), choose 'first'. If females were to be the reference category, 'last' would be chosen. Click 'Continue' to return to the Logistic Regression box.

Once back in the Logistic Regression box, select 'Options'. I have not included any of these in this example, but some analyses may call for one or more of these statistics to be included.



Once you have clicked 'OK' in the 'Logistic Regression' box, the analysis will be run, and a number of tables will be generated. However, only a small number of them are crucial in interpreting the data. The first is the model summary, which gives two statistics explaining the variation.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1234558.685[a]	.018	.028

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Of Cox & Snell R Square and Nagelkerke R Square, the latter is considered the more accurate as it is an adjusted version of the former so that 1 may be included. A value equal to or close to 1 suggests that the model is a good fit for explaining variances. As this model produces a figure of 0.28, only 3% of the variance is explained by the model.

The next table to be considered is the Classification Table. The note at the bottom tells us that if the probability of a person being diagnosed with COPD is greater than 0.5, it will be classified as a positive result. This table tells us that the model accurately predicts the outcome 79% of the time.

Classification Table[a]

Observed			Predicted		
			COPD		Percentage
			.00	1.00	Correct
Step 1	COPD	.00	975867	0	100.0
		1.00	255266	0	.0
	Overall Percentage				79.3

a. The cut value is .500

The final important table is the Variables in the Equation table. This shows the contribution of each variable within the model, and its statistical significance. The age_grp and year rows indicate that each has a significant effect (Sig [p] is less than 0.05) on the dependent variable (COPD diagnoses). The Exp(B) figures tell us how much more likely patients in each age group are to be diagnosed with COPD than the reference group (in this example, the youngest age group). The highest age group is 1.53 (i.e. 53%) times more likely to be diagnosed with COPD than the lowest.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	age_grp			18550.401	15	.000	
	age_grp(1)	-1.072	.103	107.539	1	.000	.342
	age_grp(2)	-1.245	.088	201.644	1	.000	.288
	age_grp(3)	-1.494	.078	371.278	1	.000	.224
	age_grp(4)	-1.379	.058	568.661	1	.000	.252
	age_grp(5)	-.874	.037	544.702	1	.000	.417
	age_grp(6)	-.538	.024	497.191	1	.000	.584
	age_grp(7)	-.338	.019	319.920	1	.000	.713
	age_grp(8)	-.151	.017	82.320	1	.000	.860
	age_grp(9)	.076	.015	24.547	1	.000	1.079
	age_grp(10)	.260	.015	321.200	1	.000	1.297
	age_grp(11)	.457	.014	1050.923	1	.000	1.579
	age_grp(12)	.601	.014	1859.960	1	.000	1.824
	age_grp(13)	.630	.014	2046.507	1	.000	1.878
	age_grp(14)	.618	.014	1901.025	1	.000	1.856
	age_grp(15)	.426	.015	793.427	1	.000	1.531
	year	.020	.002	157.391	1	.000	1.020
	Constant	-41.635	3.183	171.058	1	.000	.000

a. Variable(s) entered on step 1: age_grp, year.

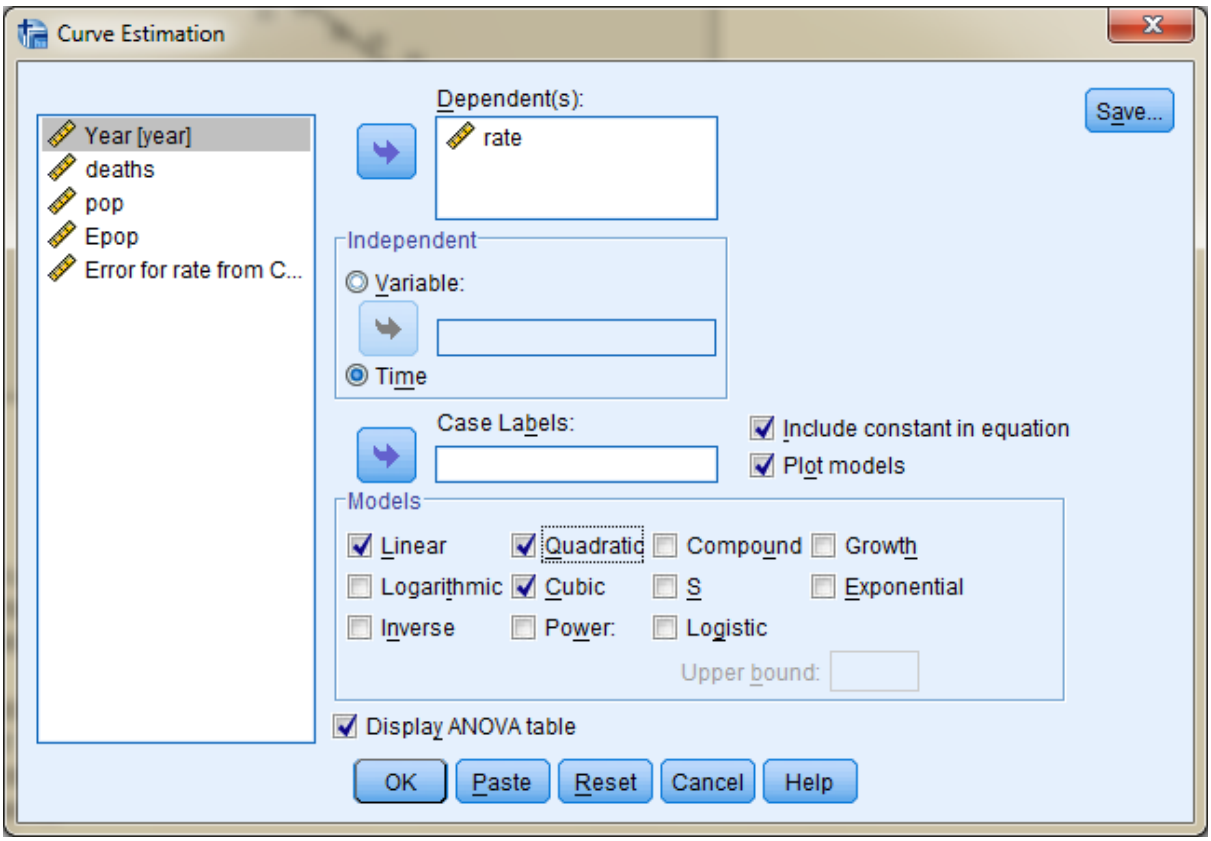
2.11: Polynomial regression

Polynomial regression can be performed in SPSS. This can be done through the menus or using syntax. Both methods are described below.

Using the menus:

Analyze → Regression → Curve Estimation

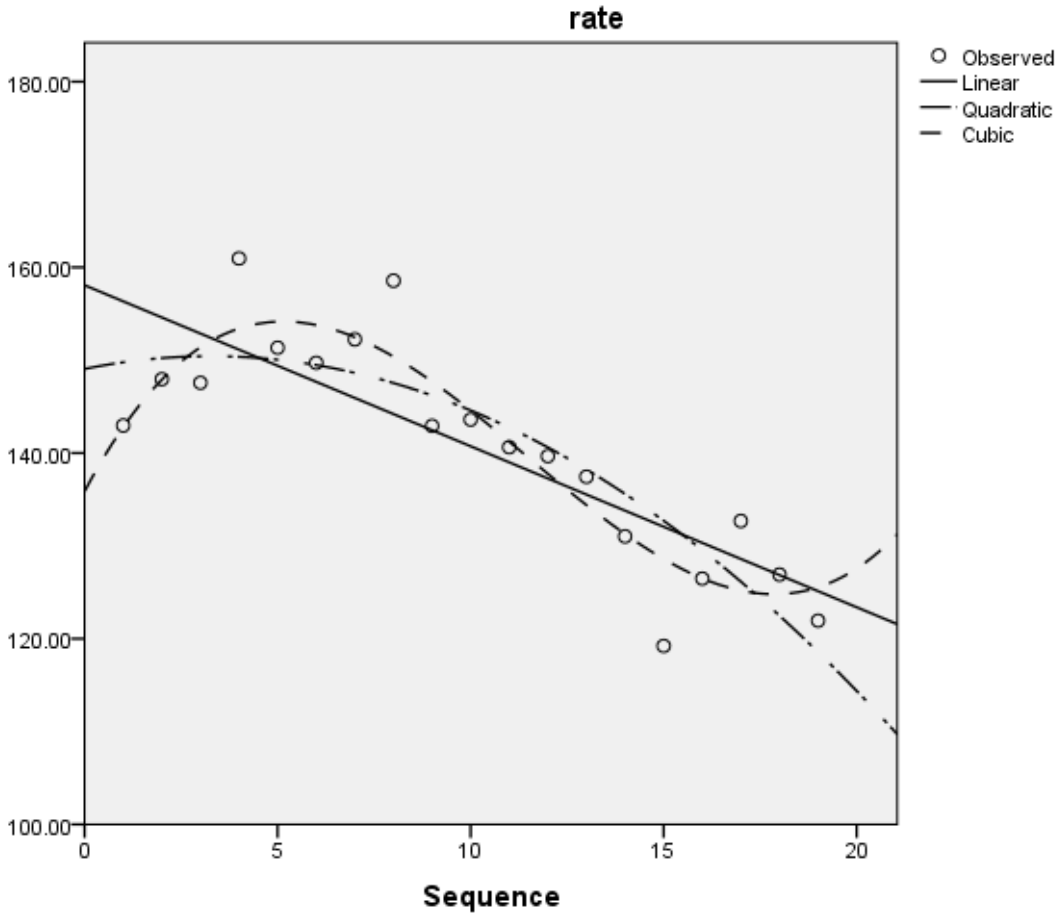
The following dialogue box will appear:



The corresponding syntax for this is:

```
* Curve Estimation.
TSET MXNEWVAR=1.
PREDICT THRU END.
CURVEFIT
  /VARIABLES=rate
  /CONSTANT
  /MODEL=CUBIC
  /PRINT ANOVA
  /PLOT FIT
  /SAVE=RESID.
```

In this example I have included linear, quadratic and cubic models. This is so that I can check which line fits my data best. Once this is identified, the statistics relating to that model are the ones that will be reported. With my data, these selections produce the scatterplot below:



The cubic model seems to fit the data quite well, although there is an unexpected curve at the right extreme.

This analysis produced the following table:

Model Summary and Parameter Estimates

Dependent Variable: rate

	Model Summary					Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.672	34.855	1	17	.000	158.072	-1.734		
Quadratic	.760	25.372	2	16	.000	149.062	.840	-.129	
Cubic	.858	30.295	3	15	.000	135.875	7.867	-.985	.029

As previously mentioned, the higher the R-squared value, the better the line fits the data. Since the cubic model produces the highest R-squared figure (0.858), it is likely that this model will be the most accurate. As an extra check, however, the residuals should be plotted. See below for details on how to calculate residuals, and create residual plots.

The desired output from these steps is normally distributed residuals (as above). Using both the R^2 figure and the residuals plot from the cubic model, we can be relatively sure that the cubic model is

the most appropriate in this instance. When the analysis is re-run, selecting only the cubic model, the following table is produced:

Model Summary and Parameter Estimates

Dependent Variable: rate

	Model Summary					Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Cubic	.858	30.295	3	15	.000	135.875	7.867	-.985	.029

The p-value (i.e. "Sig") shows that there is a strongly significant relationship, indicated by a value below 0.05, between years for age/sex standardised mortality as a result of COPD (chronic obstructive pulmonary disorder). A similar output is achieved by checking the "Display ANOVA table" box, as below:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.926	.858	.830	4.908

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	2189.212	3	729.737	30.295	.000
Residual	361.312	15	24.087		
Total	2550.524	18			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Case Sequence	7.867	2.347	3.719	3.352	.004
Case Sequence ** 2	-.985	.269	-9.586	-3.661	.002
Case Sequence ** 3	.029	.009	5.176	3.222	.006
(Constant)	135.875	5.563		24.427	.000

More information is included in the ANOVA tables, but for the purposes of this analysis either may be used. The important figures from either set of tables are b/B and the p-value.

Each B figure tells us how much mortality rate from COPD is estimated to change depending on which slope is being read. In this example, the first slope peaks at year 5 (i.e. 1999). Mortality from COPD up to 1999 is estimated to increase by 7.9. The next slope begins at the peak at 1999 and continues to the trough at around 2011. The B figure for this slope suggests that the mortality rate decreases each year by 1. From 2011 to 2014 the mortality rate is estimated to increase by 0.03 each year.

The F value is a measure of the linear regression relationship between the dependent and independent variables. The p-value in the middle table relates directly to the F value, and indicates the strength of the relationship between all variables. The p-values in the bottom table are below 0.05 for each interaction, suggesting that the difference between all years is significant.

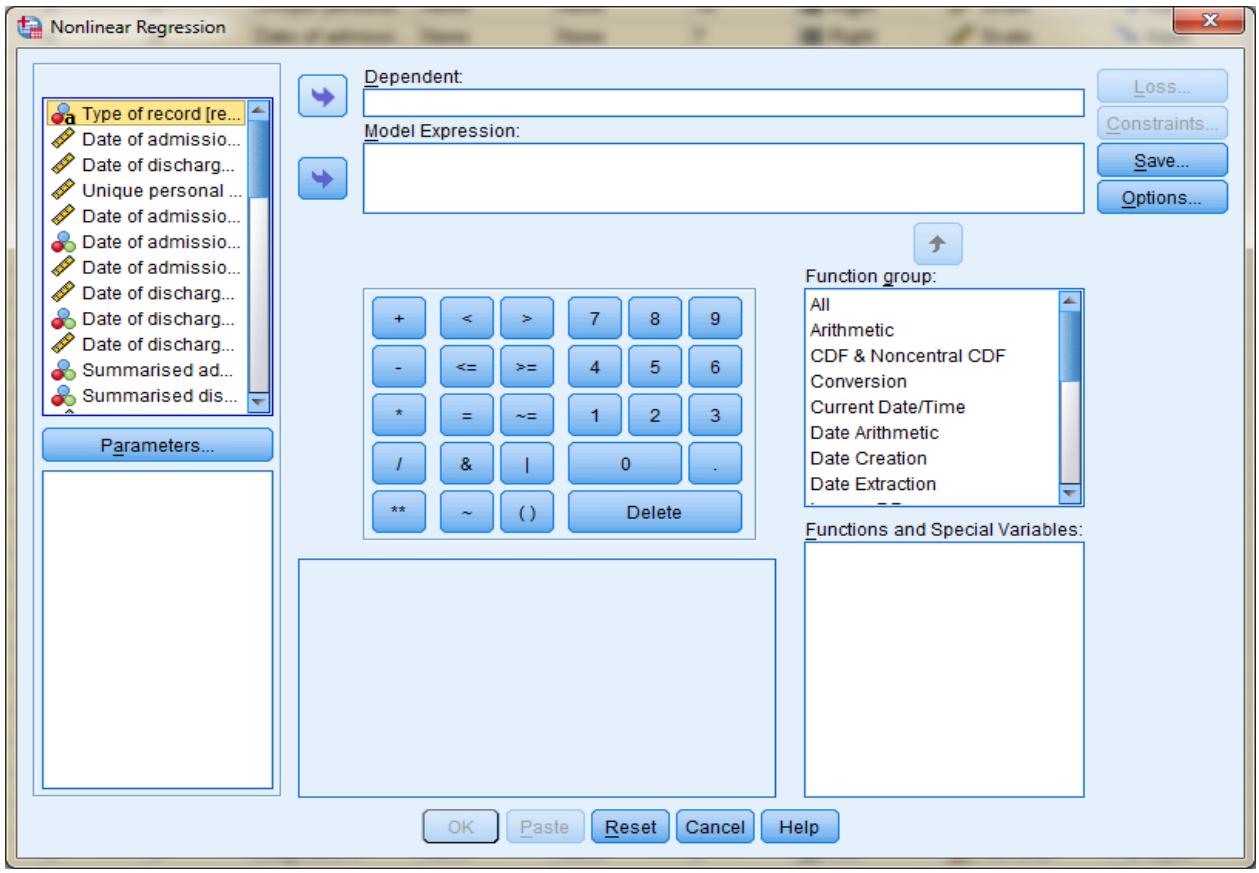
2.12: Restricted cubic splines

Harrell (2001)[26; 27] suggests the following locations (assuming the maximum is 1) for the knots, depending on how many knots the analyst would like to have.

N of knots	Quartile 1	Quartile 2	Quartile 3	Quartile 4	Quartile 5	Quartile 6	Quartile 7
3	0.1	0.5	0.9				
4	0.05	0.35	0.65	0.95			
5	0.05	0.275	0.5	0.725	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.5	0.6583	0.8167	0.975

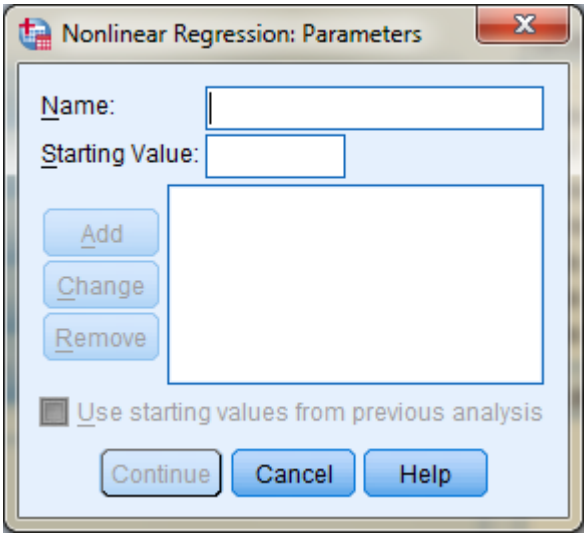
SPSS has restricted functionality when it comes to applying restricted cubic splines to data. As such, the two options available are quite complex and restrictive. It is possible to perform this method using the menus, but this requires the analyst to manually insert the equation into the dialogue box that opens through:

Analyze → Regression → Non-linear



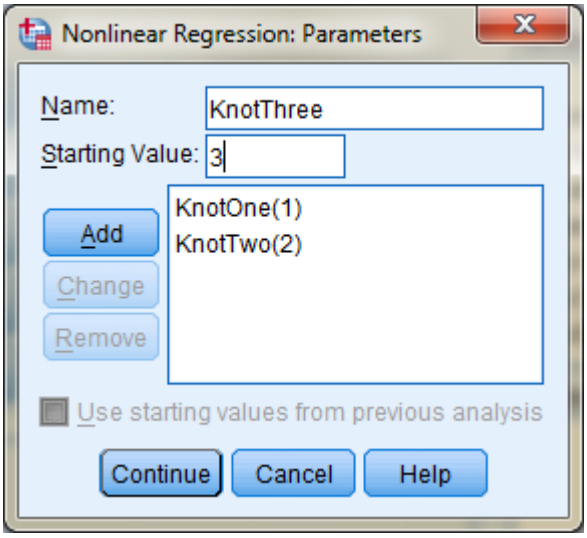
Choose your dependent variable and use the arrow to insert it into the 'dependent' box. The next step is defining parameters. Click the 'Parameters' button, for the dialogue box below:

[26] Harrell, F.E. (2001) *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* New York: Springer-Verlag New York.
[27] Table accessed at: Croxford, R. (2016) *Restricted Cubic Spline Regression: A Brief Introduction.* Available at: http://support.sas.com/resources/papers/proceedings16/5621-2016.pdf (Opens PDF; Accessed: 4 August 2016).



For each parameter (e.g. coefficients and knots), the following steps must be taken:

1. Type the name of the parameter into the box. This must not be the same as any variables in the data set.
2. Type the starting value of that parameter.
3. Click 'Add' to ensure that it's defined (see below):



Once all parameters have been defined click 'Continue' to return to the first dialogue box. You can add new parameters at any time before running the model. In the 'Model Expression' box, enter the regression equation. In this example, it is:

$$ba0 + ba1 * xa + bb1 * (xa - KnotOne) * (xa\ GE\ KnotOne) +$$

$$bc1 * (xa - KnotTwo) * (xa\ GE\ KnotTwo) +$$

$$bd1 * (xa - KnotThree) * (xa\ GE\ KnotThree)$$

Where:

- ba0 is the intercept, or where the slope crosses the Y-axis;
    - this is calculated first by calculating the slope:

$$m = r\left(\frac{S_y}{S_x}\right)$$

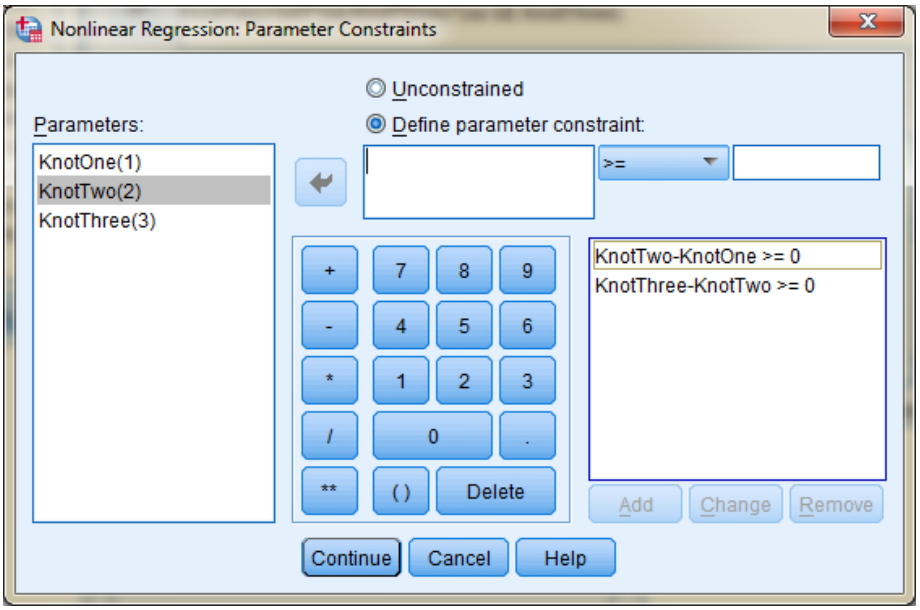
        - m denotes the slope
        - r is Pearson's *r*, the correlation between x and y
        - S_x is the standard deviation of x
        - S_y is the standard deviation of y
    - then by using the following equation to find the intercept:

$$b = \bar{y} - m\bar{x}$$

        - b is the intercept value
        - $\bar{x}$ is the mean of the values of x
        - $\bar{y}$ is the mean of the values of y
        - m is the slope that was calculated in the previous step
- ba1 is the linear coefficient of the dependent variable (i.e. Pearson's *r*)
- xa is the predictor, or independent, variable
- bb1 is the second adjustment to the slope; bc1 is the third adjustment to the slope; bd1 is the fourth adjustment to the slope

The addition of the logical expression xa GE KnotOne (etc) means that the influence of bb1 is limited to the value of the independent variable at or after the first knot. The same is true for each successive adjustment.

If it's important that KnotTwo is greater than KnotOne, click the 'Constraints' button in the first dialogue box. In the dialogue box that appears, enter an arithmetic function that meets your requirements. In this example, the constraint will be that KnotTwo minus KnotOne must be greater than or equal to 0, and KnotThree minus KnotTwo is greater than or equal to 0. This can obviously be amended to include any calculation in any group of variables.



Once you have defined all the constraints you would like the model to have, press ‘Continue’. To save the residuals for plotting later, click the ‘Save’ button in the original dialogue box and you will be asked which values you would like to keep of ‘Predicted values’, ‘Residuals’ and ‘Derivatives’. When you have built the model you would like to use, click ‘OK’.

If there are any changes you would like to make (for instance, changing the constraints equations to be ‘less than’), select ‘Paste’, which will paste the syntax into the syntax editor. It can then be amended in whatever way you wish.

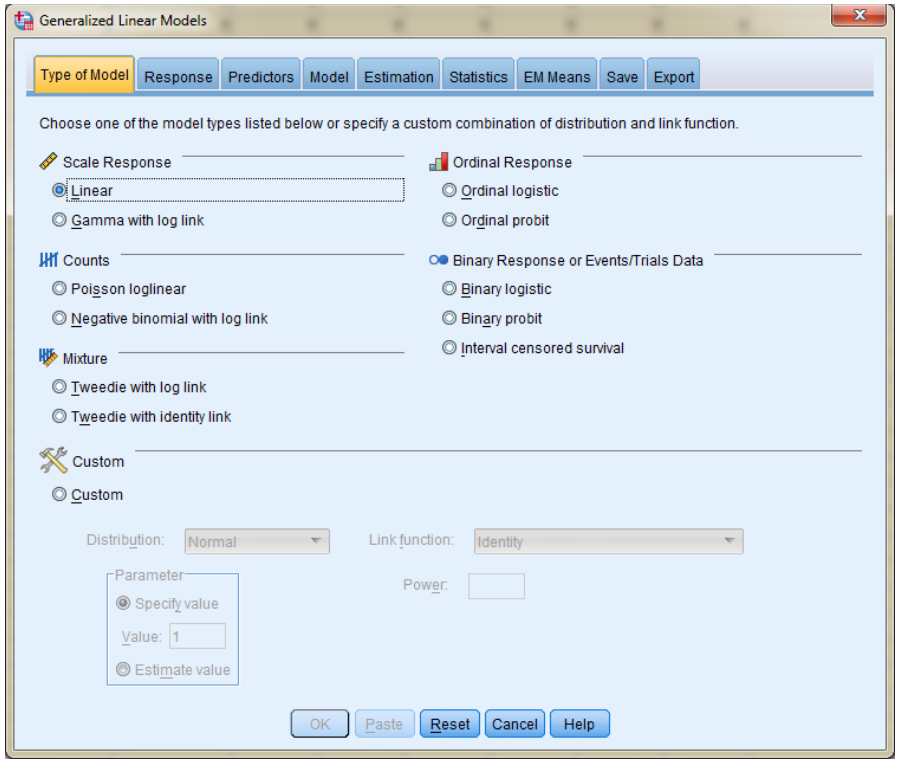
Regression analysis using restricted cubic splines

Once the knots have been defined and all data (e.g. residuals) has been included in your dataset, the derived variables may be used in regression models as the independent variables.

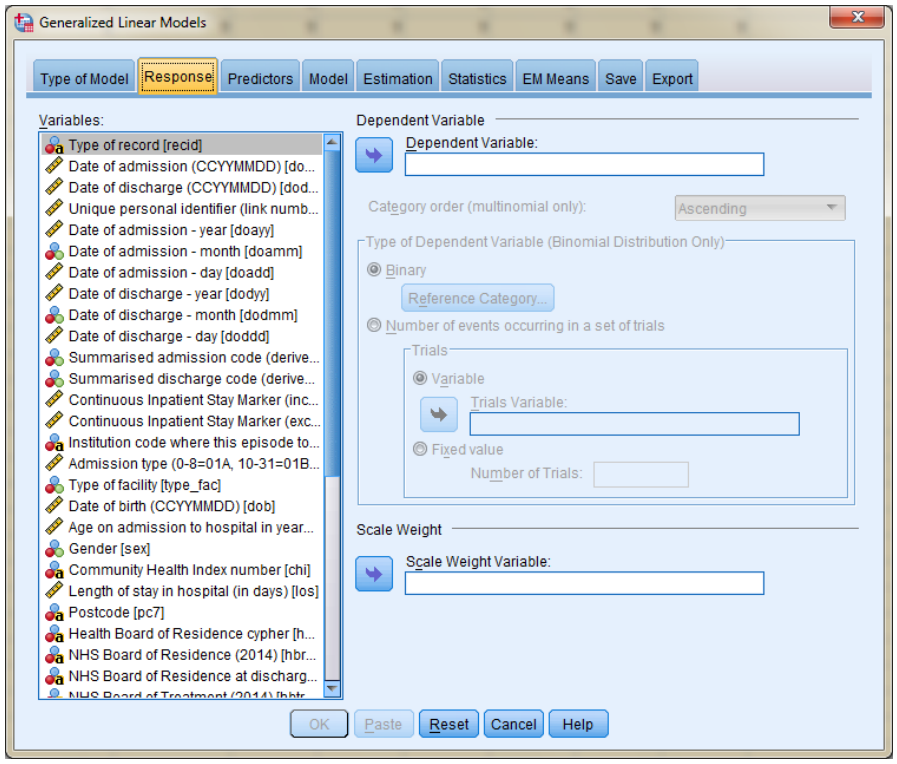
2.13: Poisson regression

As with other types of regression, this can be carried out in SPSS through the menus.

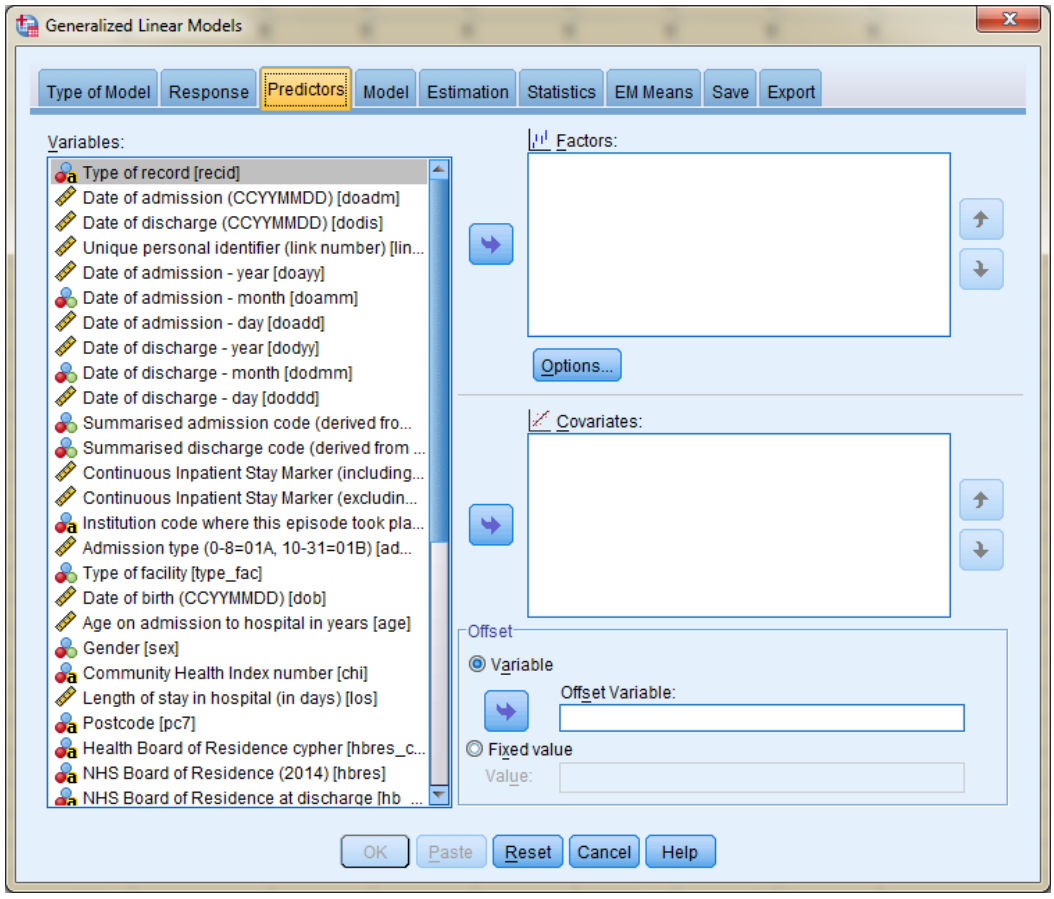
Analyze → Generalized Linear Models → Generalized Linear Models



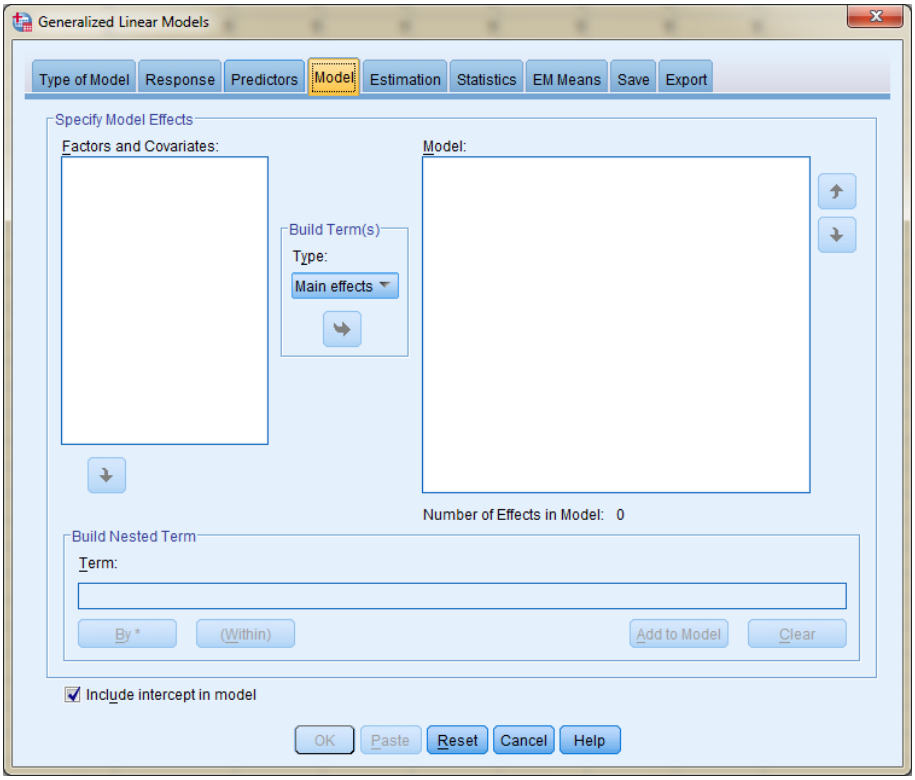
Choose 'Poisson loglinear', and then click the 'Response' tab at the top. In the box below, place your dependent variable in the 'Dependent Variable' box.



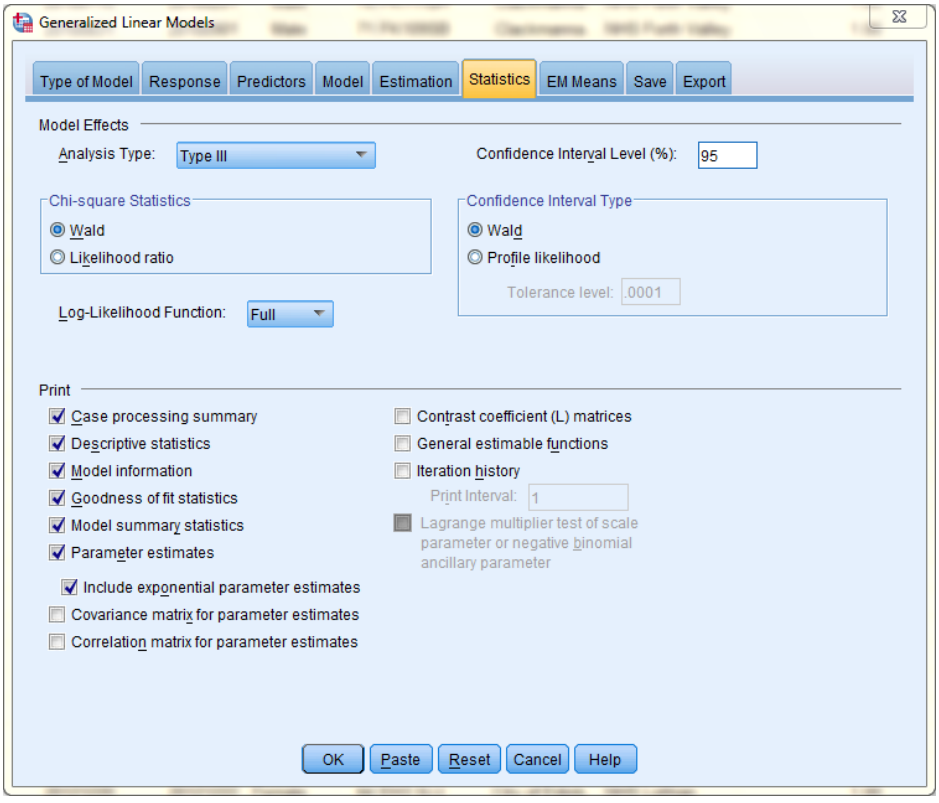
Once that has been done, choose the ‘Predictors’ tab. Categorical independent variables go into the ‘Factors’ box, while continuous independent variables are to be placed in the ‘Covariates’ box. A decision must be made about where to put them if you are analysing ordinal variables. There is no convention in SPSS to place ordinal variables in Poisson regression models.



Once you have defined your independent variables, choose the ‘Model’ tab. This is where your Poisson regression model is built. The independent variables will appear in the box on the left; to add them to the model, select them and click the right arrow. The ‘Build Term: Type’ in this example is ‘Main effects’ which for Poisson regression is correct.



Once you have entered your independent variables, click 'Statistics'. Ensure that 'Include exponential parameter estimates' is selected. Click 'OK' to run the model.



A number of tables will be produced, the most important of which are described here. This table mainly tells us if there is overdispersion between the categorical variables. In this example, we can see that there is very little difference between the number of males and females in the sample. Large

differences between groups may cause issues with the fit of the model so small variations are preferable.

Categorical Variable Information

			N	Percent
Factor	Gender	Male	616707	50.1%
		Female	614419	49.9%
		Total	1231126	100.0%

The next table tells us if overdispersion may be an issue in this model. It is important that this is not the case, as the mean and variance should be similar to each other (see section on Poisson regression).

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	COPD	1231126	.00	1.00	.2073	.40540
Covariate	Calendar Year of Discharge	1231126	2010	2014	2012.04	1.418

As we can see, the mean is 0.21. Variance is calculated by squaring the standard deviation (i.e. 0.40540^2). In this case, variance is 0.16. The ratio of mean to variance is 1.3. This means there is a small amount of overdispersion in this model. The next table provides further investigation into the difference between the mean and variance.

Goodness of Fit[a]

	Value	df	Value/df
Deviance	801797.757	1231123	.651
Scaled Deviance	801797.757	1231123	
Pearson Chi-Square	975870.915	1231123	.793
Scaled Pearson Chi-Square	975870.915	1231123	
Log Likelihood[b]	-656161.878		
Akaike's Information Criterion (AIC)	1312329.757		
Finite Sample Corrected AIC (AICC)	1312329.757		
Bayesian Information Criterion (BIC)	1312365.827		
Consistent AIC (CAIC)	1312368.827		

Dependent Variable: COPD

Model: (Intercept), sex, year

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

Many tests are performed to determine goodness of fit. However, for the purposes of this analysis, we'll look at the 'Value/df' values of the Pearson Chi-Square test. This is 0.8; a perfectly dispersed dataset will have a value of 1. This indicates that the data is slightly underdispersed. In smaller sample

sizes, this would not be an issue. In this example, our sample size is very large so it may mean that this model isn't the right one for this analysis.

The following table gives the effect on the model of the independent variables. We can see that both gender and year had a strongly significant effect on the dependent variable.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	134.011	1	.000
sex	1330.206	1	.000
year	121.347	1	.000

Dependent Variable: COPD

Model: (Intercept), sex, year

The final table gives more information about the effect of each value of the categorical independent value (gender) on the dependent variable (hospital admissions for COPD). The most informative column in this table is the 'Exp(B)' column. From this, we can say that men are less likely to be admitted to hospital for COPD than women; for every 100 women admitted to hospital with a diagnosis of COPD, almost 87 men admitted to hospital will be diagnosed with COPD. However, we can also see that there for each successive year, it is 1.016 times more likely (i.e. 1.6%) that a person will be diagnosed with COPD. These relationships are statistically significant, since 'Sig' is less than 0.05.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-32.468	2.8109	-37.977	-26.958	133.415	1	.000	7.933E-015	3.212E-017	1.959E-012
[sex=1]	-.145	.0040	-.153	-.137	1330.206	1	.000	.865	.859	.872
[sex=2]	0[a]	.	.	.	.	.	.	1	.	.
year	.015	.0014	.013	.018	121.347	1	.000	1.016	1.013	1.018
(Scale)	1[b]									

Dependent Variable: COPD

Model: (Intercept), sex, year

a. Set to zero because this parameter is redundant.

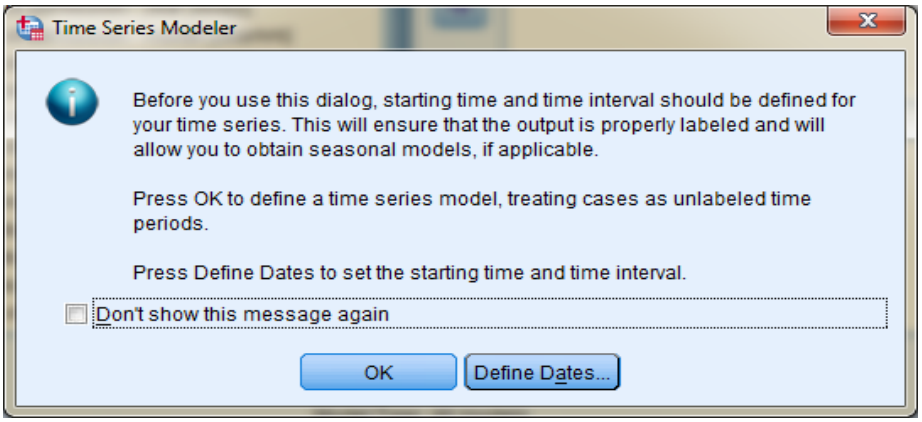
b. Fixed at the displayed value.

2.14: ARIMA

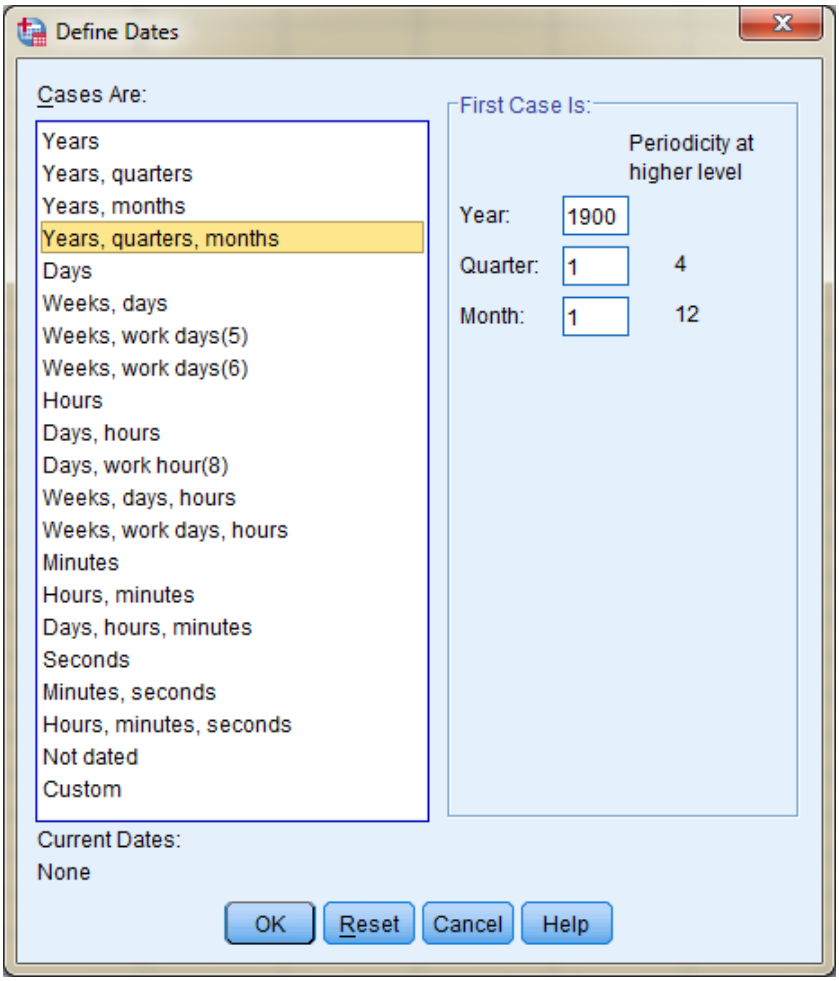
The process for this is slightly different than that of other types of regression. However, it can still be accessed through the SPSS menus.

Analyze → Forecasting → Create Models

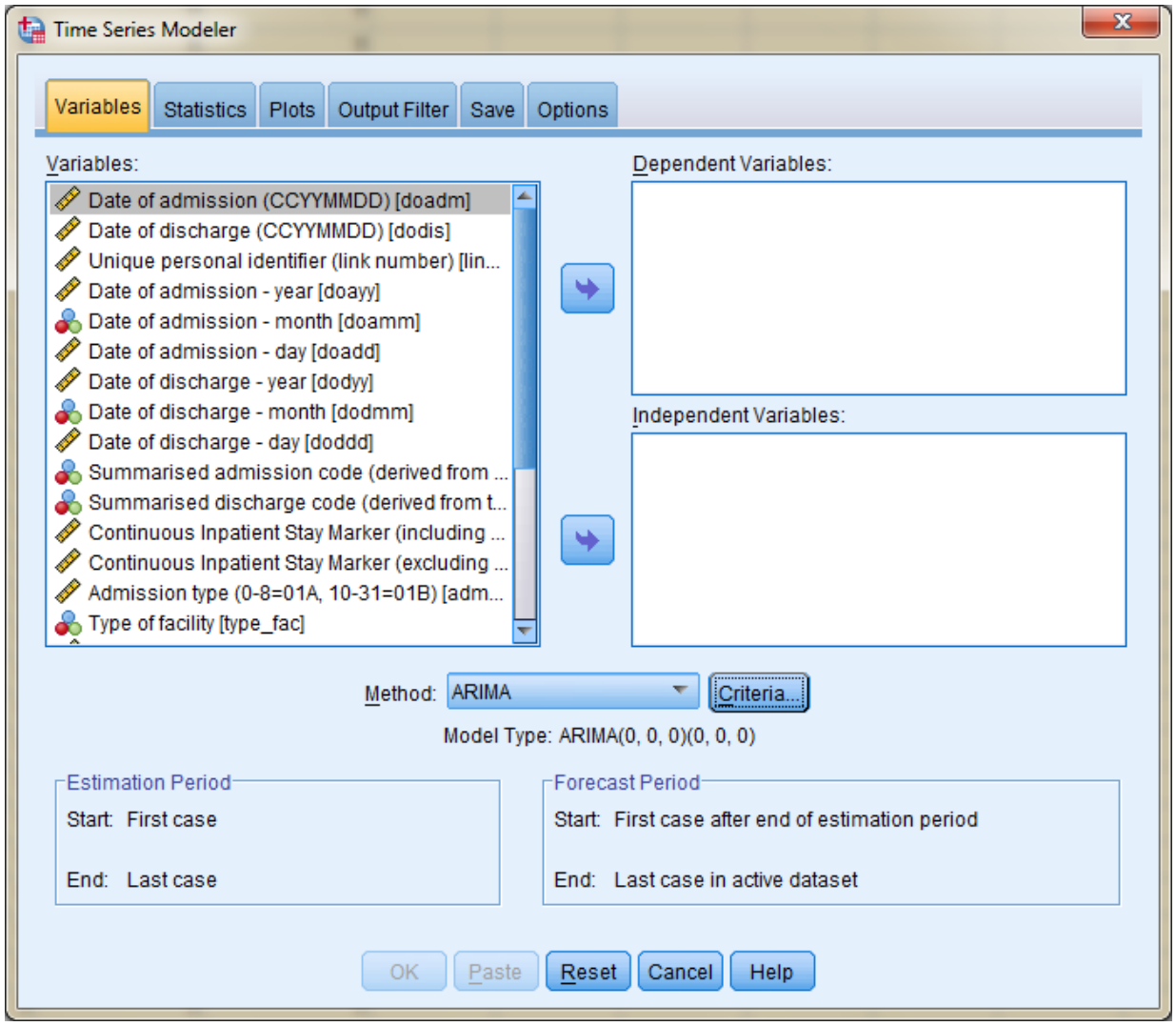
You will be presented with an information box:



If you wish to define the start and end dates of your time series, click 'Define Dates', and define the parameters.

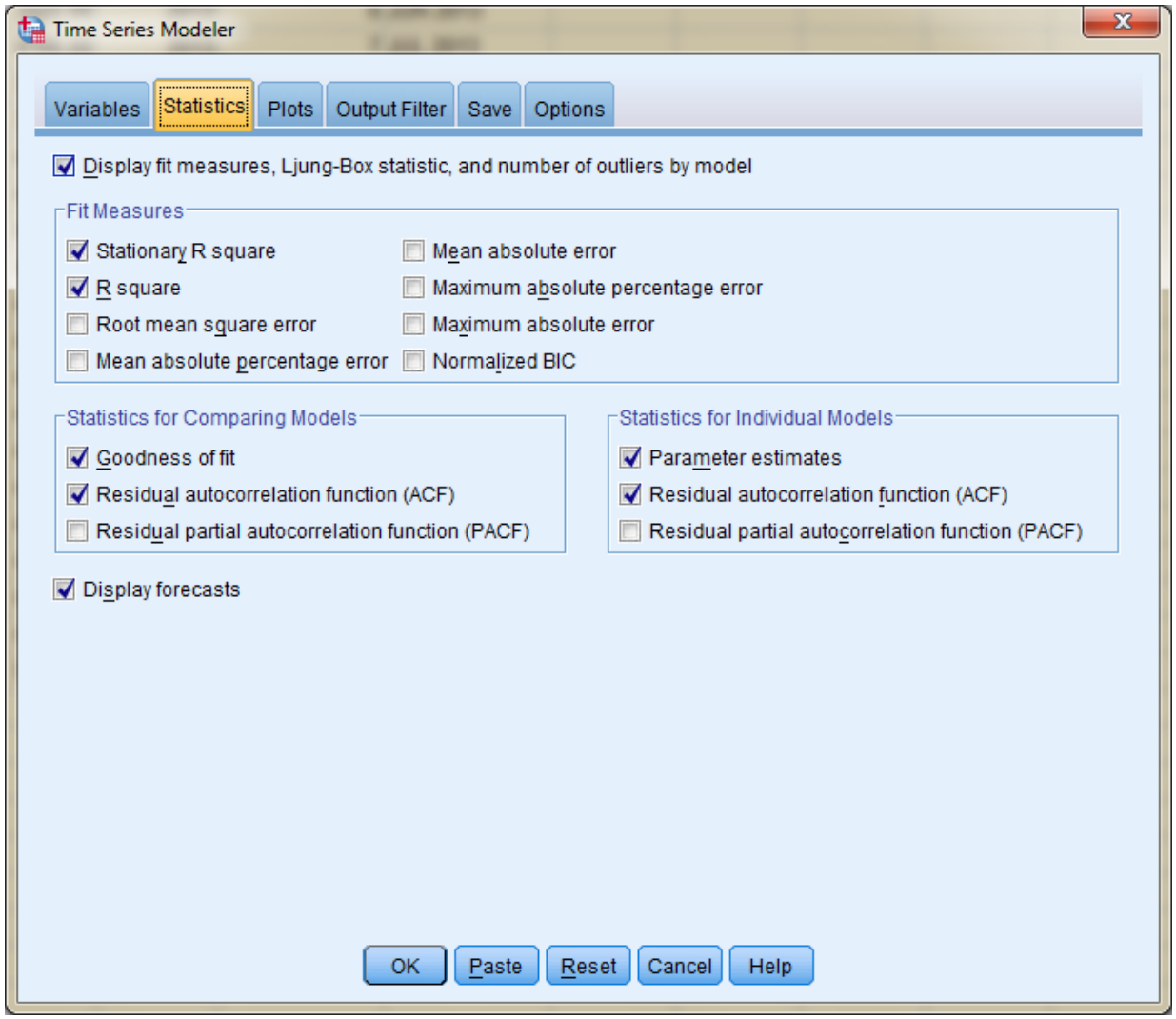


Once you have decided (or not) on the dates you would like to perform the analysis on, click ‘OK’. You will be taken to a dialogue box as below:



The default method is ‘Expert Modeler’, but that includes Exponential Smoothing in the analysis. To include only ARIMA, select it from the dropdown menu. The inclusion of independent variables is optional in this model. If independent variables are included in the model they will be treated similar to how they would be treated in a regression model. In this example I will be forecasting number of COPD admissions (dependent variable) per month and year (independent variables). Once you have placed the variables in the correct boxes go into the ‘Statistics’ tab.

Ensure that under ‘Statistics for Comparing Models’, the boxes checked above are selected before you run your analysis. Once you have done so press ‘OK’ to run the model.



The following table is produced when the analysis is run:

Model Fit

Fit Statistic	Mean	SE	Minimum	Maximum	Percentile						
					5	10	25	50	75	90	95
Stationary R-squared	.336	.	.336	.336	.336	.336	.336	.336	.336	.336	.336
R-squared	.336	.	.336	.336	.336	.336	.336	.336	.336	.336	.336
RMSE	315.945	.	315.945	315.945	315.945	315.945	315.945	315.945	315.945	315.945	315.945
MAPE	5.050	.	5.050	5.050	5.050	5.050	5.050	5.050	5.050	5.050	5.050
MaxAPE	23.504	.	23.504	23.504	23.504	23.504	23.504	23.504	23.504	23.504	23.504
MAE	219.458	.	219.458	219.458	219.458	219.458	219.458	219.458	219.458	219.458	219.458
MaxAE	1401.515	.	1401.515	1401.515	1401.515	1401.515	1401.515	1401.515	1401.515	1401.515	1401.515
Normalized BIC	11.716	.	11.716	11.716	11.716	11.716	11.716	11.716	11.716	11.716	11.716

The R-Squared tells us how well the model fits our data, as mentioned above. This analysis accounts for 33.6% of the variance seen in the data.

Most of the information you need is in the below table:

Model Statistics

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	R-squared	Statistics	DF	Sig.	
COPD-Model_1	2	.336	.336	34.973	18	.010	0

This table includes the R-squared figure, and also the significance of the relationship between the values of COPD admissions. In this case it is significant, meaning that the difference in values between each month over 5 years is significantly different from the one next to it, when autocorrelation is controlled for.

Appendix 3 Further Reading

Websites:

Andale. 2013. *F Statistic: Definition and How to find it*. Available at: http://www.statisticshowto.com/f-statistic/. [Accessed 19 December 2016].

Andale. 2016. *What is Poisson Regression?*. Available at: http://www.statisticshowto.com/poisson-regression/. [Accessed 19 December 2016].

Barry, C., Runkel, P., Rudy, K., Frost, J., Fox, G., Heckman, E. et al. 2016. *Regression Analysis,* Available at: http://blog.minitab.com/blog/regression-analysis-3. [Accessed 30 November 2016].

Bottle, A., Jarman, B., Aylin, P. 2011. *Strengths and weaknesses of hospital standardised mortality ratios*. British Medical Journal, 342, p. 7112.

Ruth Croxford, R. 2016. *Restricted Cubic Spline Regression: A Brief Introduction*. Available at: http://support.sas.com/resources/papers/proceedings16/5621-2016.pdf. [Accessed 19 December 2016].

Grace-Martin, K. (2012). *Why Use Odds Ratios in Logistic Regression.* Available at: http://www.theanalysisfactor.com/why-use-odds-ratios/. [Accessed 04 January 2017].

Harrell, F, 2001. *Regression Modeling Strategies*. 1st ed. New York: Springer-Verlag.

Information Services Division. 2016. *Publications*. Available at: http://isdscotland.org/Publications/index.asp [Accessed 10 November 2016].

Institute for Digital Research and Education UCLA. n.d. *SPSS Annotated Output: Regression Analysis,* Available at: http://www.ats.ucla.edu/stat/spss/output/reg_spss.htm. [Accessed 30 November 2016].

Institute for Digital Research and Education UCLA. n.d. *SPSS Topics: Regression,* Available at: http://stats.idre.ucla.edu/spss/output/regression-analysis/. [Accessed 30 November 2016].

Investopedia. 2016. *Autoregressive*. Available at: http://www.investopedia.com/terms/a/autoregressive.asp. [Accessed 19 December 2016].

Khan Academy. 2016. *Confidence intervals (one sample)*. Available at: https://www.khanacademy.org/math/statistics-probability/confidence-intervals-one-sample. [Accessed 19 December 2016].

Laerd Statistics. 2013. *Poisson Regression Analysis using SPSS Statistic*. Available at: https://statistics.laerd.com/spss-tutorials/poisson-regression-using-spss-statistics.php. [Accessed 30 November 2016].

Jeff Morrison, J. 2016. *Autoregressive Integrated Moving Average Models*. Available at: http://www.forecastingsolutions.com/arima.html. [Accessed 19 December 2016].

Robert Nau, R. 2016. *Regression diagnostics: testing the assumptions of linear regression*. Available at: http://people.duke.edu/~rnau/testing.htm. [Accessed 19 December 2016].

Scottish Public Health Observatory. 2016. *Online Profiles Tool*. Available at: https://scotpho.nhsnss.scot.nhs.uk/scotpho/homeAction.do [Accessed 10 November 2016].

Statistics Solution. 2016. *Logistic Regression Analysis in SPSS*. Available at: http://www.statisticssolutions.com/the-logistic-regression-analysis-in-spss/. [Accessed 10 November 2016).

Statistics Solutions. 2016. *Assumptions of Linear Regression*. Available at: http://www.statisticssolutions.com/assumptions-of-linear-regression/. [Accessed 19 December 2016].

Stone, C. J., 1986. *The Dimensionality Reduction Principle for Generalized Additive Models*. The Annals of Statistics, 14(2), 590-606. Available at: http://www.jstor.org/stable/2241237 [Accessed 19 December 2016].

Szumilas, M. (2010). *Explaining Odds Ratios*. Journal of the Canadian Academy of Child and Adolescent Psychiatry, 19(3), pp. 227-229.

Wagner, A.K., S.B., Soumerai, F., Zhang, D., Ross-Degnan, 2002. *Segmented regression analysis of interrupted time series studies in medication use research*. Journal of Clinical Pharmacy and Therapeuticals, 27(4), 299-309.

Wikimedia Foundation. 2016. *Outliers.* Available at: https://en.wikipedia.org/wiki/Outlier#Detection. [Accessed 19 December 2016].

Wikimedia Foundation. 2016. *p-value.* Available at: https://en.wikipedia.org/wiki/P-value. [Accessed 3 March 2017].

Wikimedia Foundation. 2016. *Regression Analysis.* Available at: https://en.wikipedia.org/wiki/Regression_analysis. [Accessed 9 November 2016].

Wuensch, K. L. (2016). *Binary Logistic Regression with SPSS*. Available at: http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.pdf. [Accessed 04 January 2017].